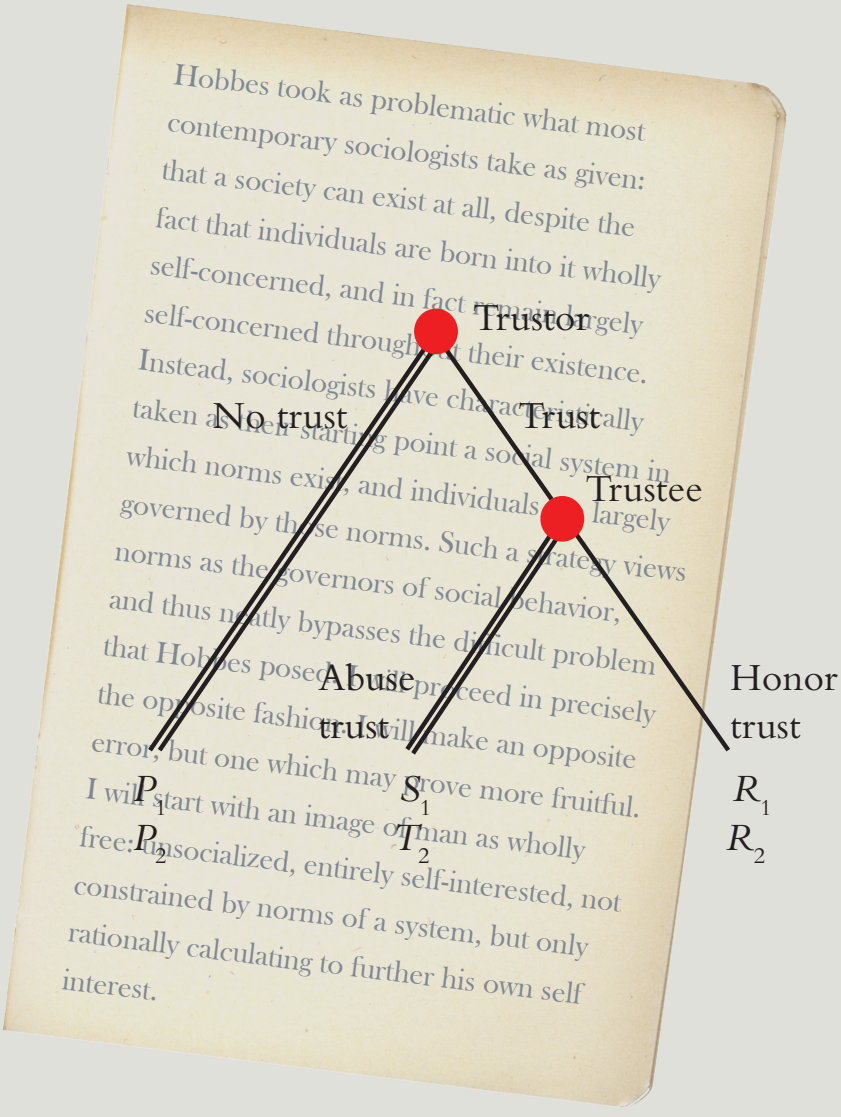
Rational Models

Werner Raub



Universiteit Utrecht

Rational Models

Rational Models

Expanded version of farewell lecture
as Dean of Social and Behavioral Sciences
at Utrecht University

Werner Raub

College ter gelegenheid van het afscheid
als decaan van de faculteit Sociale Wetenschappen
aan de Universiteit Utrecht
in verkorte vorm uitgesproken op maandag 18 september 2017

Raub, Werner
Rational Models
Utrecht: Universiteit Utrecht
ISBN 978-90393-6804-6

Copyright © 2017 Werner Raub

Contact: w.raub@uu.nl

Jacket quotes: Coleman (1964: 166–167) and Popper ([1934] 1973: 7–8)

Layout: Monique Janssen-van Rooijen, Universiteit Utrecht | Renate Siebes, Proefschrift.nu

Cover design: Monique Janssen-van Rooijen, Universiteit Utrecht

Voor Reinhard Wippler –
leermeester en model

Contents

Meneer de Rector,
Ladies and Gentlemen,

This lecture is about "rational models" in three respects. First, rational models in the intuitive sense of being fit for purpose: models that are an appropriate means to a given end – they contribute to the growth of knowledge in sociology. I will sketch core characteristics of rigorous sociology that I consider to be fit for purpose. Second, rational models in the technical sense of using assumptions about rational behavior for sociological theories. More specifically, I will sketch rational choice models of trust and, more generally, cooperation in social dilemmas as well as empirical research and results of such research in this field.

I will look at rational models in the first and the second sense through the lens of what I learned throughout my academic career and with a focus on new results obtained during that career. I will be looking back. I will ask "What has been done?" and will try to provide answers. I will resist the temptation of telling others how to proceed. I will seldom be looking forward and ask "What remains to be done?" Still, I find it a pleasant prospect to be able to engage in new research in the upcoming years, and more so than in the recent past. Hopefully I will be able to contribute to some research that remains to be done. Looking back provides ample opportunity to refer rather shamelessly to and make use of my earlier work and the work of close collaborators. I will readily yield to that temptation.[1] Much of the relevant literature is rather technical in nature. I will try to avoid "technicalities" as much as possible and to sketch basic ideas so that they are accessible to non-experts.

The lecture marks my retirement as Dean of Social and Behavioral Sciences. I served as Dean for more than five years. I had been Vice-Dean before, after some fifteen years as leader of the Utrecht Department of Sociology and, together with colleagues, of our research school ICS, the *Interuniversity Center for Social Science Theory and Methodology*. Quite a long period spent in different roles that offered many opportunities to have a go at applying sociological knowledge in running a

---

Nuffield College of the University of Oxford provided hospitality while I was preparing this lecture. I wrote the text at Nuffield in January 2017. Initially meant to be a 45-minutes lecture, the manuscript kept growing – *het liep uit de hand*, as they say in Dutch – and became a booklet. Enjoying a position with fewer incentives for "salami slicing" publication behavior, I decided to leave it at that and to postpone preparing the "real" lecture until the 2017 summer vacation period. I am indebted to Nuffield College for my Senior Research Fellowship, which allowed me to spend time at Nuffield each year in a convenient and pleasant academic environment. Since 2011, this has been a welcome opportunity for research and writing. Madelon Pieper and Monique Janssen-van Rooijen prepared the booklet for publication.

[1] Likewise, I will not hesitate to refer to literature from quite a while ago that has been influential in the formative phase of my academic career and that I still consider to be worth the effort.

Department, an interuniversity research school, and a Faculty. Hence, I include brief sketches of how rational models can be used to generate ideas on how to run such institutions in a rational manner, namely, so that they provide contexts that are fit for purpose for scholars and for students: contexts providing good conditions for teaching and learning as well as for research. This is the third respect in which my lecture is about "rational models."

Die Affen als Kunstrichter

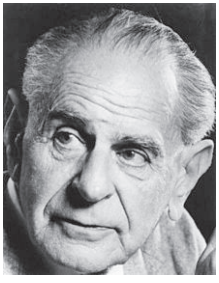
More visually, one would like to minimize the risk of the team running a Department, a research school or the Executive Board of a Faculty coming to resemble the apes in the painting *Monkeys as Judges of Art* (*Die Affen als Kunstrichter*) by Gabriel Cornelius von Max. Rather, one would like to increase the likelihood of such a team resembling Rembrandt's *Sampling Officials* (*De Staalmeesters*).

De Staalmeesters

Providing rational models in this third sense is useful too, because it highlights practical applications of abstract theory. In an empirical science such as sociology, abstract theory should not be *l'art pour l'art* but should be able to guide empirical research as well as contribute to sensible policy making. Moreover, focusing on some general principles of running academic institutions avoids the risk of hampering those who are now in charge by interfering in how they deal with specific issues on their current agenda.

1. Rigorous sociology[2]

"Rigorous sociology" is a useful label for a family of research programs that conceive of sociology as a problem- and theory-guided discipline. Theory construction aims to explain social phenomena. Competing explanations and theoretical pluralism contribute to the growth of knowledge. Explanations involve deductive arguments or variants of such arguments. Therefore, theory construction involves more than specifying sets of hypotheses. Rather, theory construction involves specifying assumptions, including but not limited to hypotheses, as well as specifying implications of these assumptions. Due to the focus on implications, analytical rigor is an important ingredient. Implications should include *testable* implications: empirical content in the sense of testability, at least "in principle," is a criterion for appraising sociological theories. Likewise, empirical tests of implications are a core aim and empirical corroboration is a criterion for appraising theories in addition to testability. Therefore, linking theory construction with empirical research and statistical modeling becomes an aim, too. This is sociology in the spirit of Popper's ([1934] 1973, 1963, 1972) philosophy of science, broadly conceived.



Karl Popper

Micro-macro links, Coleman's diagram, and an example

Much of rigorous sociology is about *micro-macro links* because it attempts to explain phenomena at the level of social systems as well as at the level of individual behavior and because it focuses on how the social system level and individual behavior are related. "Macro" refers to social systems such as a family, a team of researchers, a business firm, a university, or a society, whereas "micro" refers to individuals (see Coleman 1986a: 346).[3] The macro-level thus refers to phenomena that are described by concepts referring to properties of social systems. In terms of size, "macro" may refer not only to large but also to small social systems. The system might be a network involving a sizeable number of actors but it might also be, for example, a dyad, a triad, or a small group. The micro-level refers to properties of individuals, such as their preferences, their information, and their behavior.

---

[2] Parts of this chapter draw on materials from Raub, Buskens, and Van Assen (2011) as well as Raub and Voss (2017).

[3] There are also well-known examples with "corporate actors" such as organizations (Coleman 1990: Part III and IV) on the micro-level.

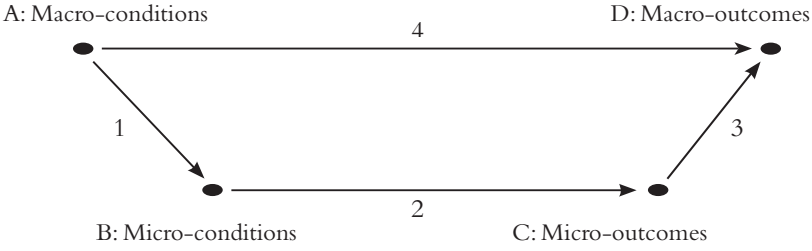
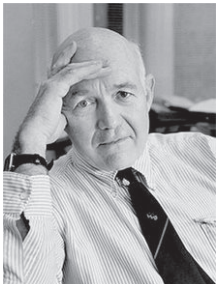


Figure 1: Coleman's diagram.



James Coleman

James Coleman was an outstanding sociologist of the twentieth century. We are indebted to him for his seminal contributions to diverse fields of the discipline. His diagram depicting micro-macro links (e.g., Coleman 1987a and Coleman 1990: Chapter 1) has become prominent and is a useful tool for clarifying the relevant issues. Let us consider his diagram together with an example, namely, co-authors working on a joint paper. When thinking through this example, keep analogies in mind such as R&D-alliances involving firms and possibly also universities or teachers who are co-teaching a course.

The set of co-authors is a small social system at the macro-level and we might want to explain why they succeeded or failed in writing a high-quality paper, our macro-outcome, indicated by Node D in the diagram. The likelihood of the macro-outcome "high-quality paper" will depend, among other things, on the time and effort each of the co-authors invests in the common project. These individual investments are micro-outcomes in terms of behavior at the micro-level, indicated by Node C. Arrow 3 in the diagram represents assumptions about how actors' behavior affects the macro-outcome. I follow Lindenberg (1977) and use "transformation rules" as a label for such assumptions about micro-to-macro relations. Explaining the individual investments of the co-authors may also help explaining the macro-outcome "success" or "failure" of the joint paper.

In addition to working on the joint paper, each of the co-authors will also be involved in other projects as well as in teaching. Since resources like time and energy are limited, each co-author must therefore choose how much of the stock of individual resources to allocate to the joint paper versus to other projects and obligations. Let us assume that individual investments will depend, among other

things, on the value one attaches to a high-quality joint paper compared to, say, the quality of another individual paper in progress and the quality of a course one must teach. Let us also assume that individual investments of time and effort in the joint paper will depend on how much one believes the other co-authors will invest. After all, the quality of the joint paper depends on their investments, too. For our example, Node B in Coleman's diagram then represents assumptions about the behavioral alternatives of a co-author, assumptions about the values a co-author attaches to different outcomes of investment decisions, and assumptions about his[4] beliefs concerning the behavior of his colleagues. In general, Node B represents propositions describing micro-conditions. These propositions refer to "independent variables" in assumptions about regularities of individual behavior or, more ambitiously, in a theory of individual behavior. Arrow 2 represents the micro-theory itself, that is, a set of general assumptions about how micro-conditions affect micro-outcomes.

By now, you may already intuit that the co-authors in our example could be facing a dilemma. It may be that each prefers the outcome in which everybody invests his fair share of time and energy in the joint project, with a high-quality paper as an outcome, to the alternative outcome of low investments and the failure of the joint project. Still, given that the quality of the joint paper depends not only on his own investments but also on those of the colleagues involved, each co-author may have an incentive to divert own investments to other individual projects or to teaching obligations. After all, a co-author may be hoping for a high-quality paper even if his own investments are limited, thanks to efforts of his colleagues. Or, a co-author may fear that colleagues will not invest their fair share and that a high-quality paper will be infeasible anyway.

In sociology, we are not only interested in explaining micro- and, particularly, macro-outcomes, we are also interested in how macro-conditions affect micro- and macro-outcomes. Macro-conditions are represented by Node A in Coleman's diagram. In our example, such macro-conditions might be that the co-authors are tenured members of the same Department, making it likely that they have been involved in other joint projects before, will be involved in new joint projects in the future and, moreover, are part of a dense network of colleagues, including colleagues not involved as co-authors of the joint paper. Such macro-conditions may affect micro-conditions. For example, experiences from previous joint projects and information received from other colleagues in the Department may affect a co-author's beliefs about the likeliness of another co-author investing his fair share. Also, each co-author's investments in the current joint project may have implications for the behavior of colleagues in future joint projects and even for the future behavior

[4] "He" and "his" are used to facilitate readability and without intending gender-bias.

of other colleagues in the Department. Moreover, co-authors may now have new and more complex options such as making their behavior conditional on the previous behavior of colleagues. Arrow 1 represents assumptions about how macro-conditions affect micro-conditions. Again, I follow Lindenberg (1981) and label them "bridge assumptions." Finally, Arrow 4 in the diagram represents propositions about empirical regularities at the macro-level, say, an association between macro-conditions and macro-outcomes.

Note that "micro-macro" is ambiguous from the perspective of Coleman's diagram. In a narrow sense, "micro-macro" can refer exclusively to Arrow 3. In a broader sense, "micro-macro" can refer to explaining macro-outcomes (Node D) and macro-regularities (Arrow 4) using assumptions about individual behavior (Node B, Arrow 2), macro-conditions (Node A), as well as bridge assumptions (Arrow 1) and transformation rules (Arrow 3). I use "micro-macro" in this broader sense. Hence, I avoid cumbersome terminology like "macro-micro-macro" and systematically refer to assumptions represented by Arrow 3 as "transformation rules."

Key features of rigorous sociology focusing on micro-macro links

Without claiming completeness, I summarize five key features of rigorous sociology. First, Coleman's diagram shows that explanations of micro-outcomes include assumptions about individuals and individual behavior (Node B, Arrow 2), macro-conditions (Node A), and bridge assumptions (Arrow 1). Explanations of macro-outcomes (Node D) and macro-regularities (Arrow 4) include assumptions about individuals and their behavior (Node B, Arrow 2), macro-conditions (Node A), as well as bridge assumptions (Arrow 1) and transformation rules (Arrow 3). The diagram clearly indicates that sociological explanations include micro- as well as macro-phenomena and try to highlight macro-conditions rather than focus exclusively on micro-conditions as part of the explanation. Thus, such explanations follow the "minimal program of sociology" (Lindenberg 1977) that can be traced back to Durkheim ([1895] 1982) in his *Rules of Sociological Method*: social facts should be causally explained by other social facts.

Second, the theoretical core of the explanation of macro-outcomes or macro-regularities involves micro-level assumptions, together with bridge assumptions and transformation rules. Various arguments have been given for why using the micro-level to explain macro-phenomena is preferable to an approach that attempts to explain them exclusively in terms of macro-assumptions (see Coleman 1990: 3–5). For example, Coleman (1990: 3–4) as well as Wippler and Lindenberg (1987: 138) reason that, compared to assumptions about macro-level regularities, assumptions about regularities of individual behavior are less subject to changing boundary

conditions that affect whether or not these assumptions apply in a given situation: human nature is relatively stable in the sense that actors behave similarly under the same conditions, while associations between macro-conditions and macro-outcomes are less stable.

Third, explanations in line with Coleman's diagram follow simple principles of model building. Model building faces the trade-off between, on the one hand, simplifying assumptions that preserve tractability and analytical power allowing for the derivation of implications, including testable implications, at the cost of being less realistic, and, on the other hand, more complex and realistic assumptions that make it more difficult to derive implications. It therefore makes sense to start with a model that is as simple as possible, making simplifying assumptions explicit. Subsequently, one can introduce more complex assumptions incrementally when simplifying assumptions turn out to be problematic because, for example, implications depend heavily on such assumptions rather than being robust, or because implications fare badly in the light of empirical evidence. This procedure is known as the method of decreasing abstraction (Lindenberg 1992). The principle of sufficient complexity (Lindenberg 2001) complements the method of decreasing abstraction. It requires that even the simplest model assumptions should be complex enough to allow us to describe the phenomenon to be explained rather than simply assuming it away. In our example of the co-authors, this requires the model to at least include explicit assumptions about interdependence between the actors, in the sense that the outcomes of each co-author's behavior in terms of the quality of the joint paper depend not only on his own behavior but also on the behavior of the other co-authors.

Fourth, micro-macro models typically try to employ comparatively simple micro-level assumptions, while simultaneously trying to incorporate more complex assumptions about macro-conditions as well as specifying the transformation rules as carefully as possible (see Coleman 1987b for a succinct statement). The motivation is threefold. Micro-macro models aim to explain macro-outcomes and to incorporate macro-conditions into the explanation rather than explaining individual behavior as such. Hence, it is reasonable to allow for complexity of macro-assumptions. Furthermore, since deriving macro-implications from micro-assumptions, bridge assumptions, and transformation rules is often a non-trivial task, it is advisable to keep the micro-assumptions simple with an eye on the tractability of the model. Finally, Coleman argues that the careful specification of transformation rules is not only a core task of sociology but that sociological explanations are also often deficient precisely with respect to the transformation rules. Hence, Coleman assumes that investments in improving transformation rules will be more beneficial for theory

development in sociology than improving micro-assumptions.[5] A similar motivation for the careful specification of bridge-assumptions is straightforward.

Fifth and last, macro-outcomes are typically the result of interdependence between actors, for example in the sense that the outcomes of an actor's behavior depend not only on his own choices and possibly chance events but also on the behavior of other actors. Moreover, due to interdependence, macro-outcomes are often unintended consequences of individual behavior: the very fact that outcomes also depend on the behavior of others means that an actor's intentions need not coincide with the outcomes of his behavior.

It should be added that Coleman's diagram provides a highly stylized and simplified representation of full-fledged micro-macro models, leaving complex issues implicit. The nodes and arrows summarize possibly numerous and rather different kinds of assumptions. Also, model building involves not only the careful specification of assumptions but also, crucially, deriving implications from assumptions.

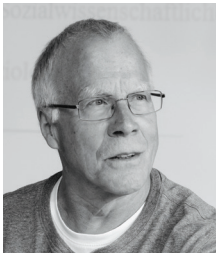
A family of research programs

Rigorous sociology as conceived here can be traced back to the Scottish Moralists of the eighteenth century (Hume, Adam Smith, Ferguson; see Schneider 1967). For example, the Scottish Moralists focused on explaining macro-phenomena, using assumptions about human nature and pushing the idea that macro-phenomena are often unintended results of individual behavior in situations with interdependence between actors. Smith's "invisible hand" is a well-known metaphor for unintended consequences of individual behavior in market contexts. Ferguson tried to account for the emergence and dynamics of norms and institutions as an unintended result of the behavior of interdependent actors. The research program of the Scottish Moralists advocates the methodological unity of the social sciences. It is therefore not surprising that it has been influential throughout the social sciences. Key ideas underlying micro-macro models can be identified in various strands of methodological individualism (see O'Neill 1973 and Udehn 2001). Research programs in economics have contributed to the development of micro-macro models and employ such models. Neoclassical economics is an exemplary case, together with, for example, Becker's (1976) "economic approach to human behavior," the new institutional economics and evolutionary economics as well as applications of

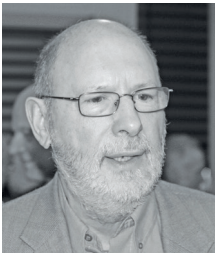
[5] Note the similarity to Coleman's earlier arguments for "synthetic theories" in his *Introduction to Mathematical Sociology* (1964a: 34ff; see also 516ff). In fact, synthetic theories can be conceived as transformation rules in micro-macro models: "it is characteristic of many of these theories that they begin with postulates on the individual level and end with deductions on the group level" (Coleman 1964a: 41).

game theory in economics, including behavioral and experimental game theory (see a microeconomics textbook like Mas-Colell, Whinston, and Green 1995 and Camerer 2003). Related research programs can be found in political science (see Riker and Ordeshook 1973 for an early textbook).

Social exchange theory (Homans 1958; Blau 1964) is commonly (e.g., Coleman 1986b) seen as the pioneer of micro-macro models in modern sociology. In Europe, the structural-individualistic research program, sometimes labeled "explanatory sociology," has developed since roughly the 1970s, pushed by authors such as Albert, Boudon, Opp, Hummell, Lindenberg, Wippler, Esser, and others (Wippler 1978 provides a succinct summary of the program; see Raub and Voss 1981 for a more detailed overview). I was lucky enough to have become acquainted with and to have been trained in rigorous sociology by teachers such as Hartmut Esser, Hans Hummell, and Reinhard Wippler, all of them important contributors to structural individualism. And I was lucky enough, too, to have been able to contribute to the development of this program together with friends and co-authors such as Thomas Voss, Jeroen Weesie, and Vincent Buskens.



Hartmut Esser



Hans Hummell

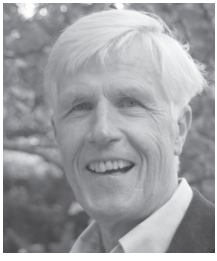
Reinhard Wippler

Thomas Voss

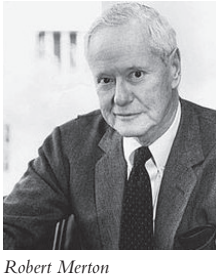
Jeroen Weesie

Vincent Buskens

Vanberg (1975) and Bohnen (1975) produced detailed studies on the roots of structural individualism in the work of the Scottish Moralists, the Austrian School of economics (including Menger, Schumpeter, Mises, and Hayek), the work of Max Weber, and methodological individualism. Lindenberg (1977, 1981; Wippler and Lindenberg 1987) and Hummell and Opp (1971; see also Opp 1979) developed the methodological tools needed for micro-macro modeling, including the use of micro-assumptions as well as bridge assumptions and transformation rules in such models. Ziegler (1972) and Hummell (1973) clarified the use of formal modeling in establishing micro-macro links. Seminal theoretical studies and many examples of the program's application can be found, for example, in Boudon (1974, 1977, 1979).


Siegwart Lindenberg

While Coleman's diagram has meanwhile become a standard way of representing micro-macro links, it is worth noting that the diagram itself has several predecessors, particularly in structural individualism (see Raub and Voss 2017 for details). Quite some years before Coleman pushed the idea, Lindenberg (1977; see also Wippler and Lindenberg 1987) developed a model comprising exactly the components of Coleman's diagram, albeit organized in a somewhat different way. Boudon (1979: Chapter V) offered a similar model and likewise made it clear that to account for processes over time, one needs to extend Coleman's diagram. More precisely, one could conceive of the macro-outcomes (node D) as the "initial" node of a subsequent micro-macro diagram and so forth, leading to a sequence of "connected" diagrams as shown in Figure 1 that account for the development of a social process over a number of periods 1, 2,… Such an extended version of the diagram also indicates how micro-macro models can endogenize the macro-conditions of a micro-macro model.


Robert Merton

It is useful to note that micro-macro models likewise resemble middle-range theories in the sense of Merton (1968). As Hedström and Udehn (2009) have shown, paradigmatic examples of middle-range theories à la Merton such as his analyses of unanticipated consequences, self-fulfilling prophecies, and the Matthew effect, typically comprise explanations of macro-phenomena in terms of macro-to-micro links represented by bridge assumptions, micro-assumptions about behavioral regularities, and micro-to-macro links through transformation rules (see already Stinchcombe 1975). "Analytical sociology" (Hedström 2005; Hedström and

Bearman 2009) is a more recent approach towards rigorous micro-macro modeling, often using agent-based computational modeling (e.g., Macy and Flache 2009) as a tool. Goldthorpe's (2016) "sociology as a population science" that aims to explain macro-level regularities likewise employs roughly the same explanatory strategy as depicted in Coleman's diagram (see Billari 2015 for an approach in demography that is similar to Goldthorpe's).

Rigorous sociology and rational choice

Sociological explanations in the spirit of Coleman's diagram employ micro-theories of behavior as the theoretical core. The diagram as such is in principle compatible with a broad range of such theories. Coleman (1986b, 1990: Chapter 1) himself, however, used his diagram as a stylized representation of the "logic" of "purposive action explanations." Such explanations use micro-theories relying on the intuitive idea that behavior is goal-oriented and incentive-driven (Harsanyi 1976: Chapter 6 and 1977: Chapter 1). Again, there is a broad range of different theories that are consistent with this idea.[6]

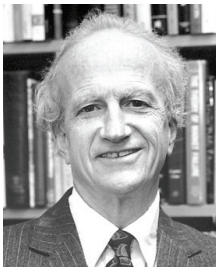
Theories of rational choice are one approach, alongside others, towards specifying what is meant by "goal-directed" and "incentive-driven" behavior (see, for example, Harsanyi 1977 for a clear and systematic exposition and Raub 1984: Chapter 2 for a summary). The core of these theories comprises a "primary definition" (Harsanyi 1977: 10–11) of rational behavior in terms of a small number of rationality postulates. Roughly, these postulates require consistency of an actor's behavior. An example of a rationality postulate is the assumption that an actor's preferences are transitive: if he prefers *A* over *B* and *B* over *C*, he likewise prefers *A* over *C*. One then shows that the primary definition implies that a utility function exists and that behavior in accordance with the rationality postulates can likewise be characterized as maximizing the utility function. These implications constitute a "secondary definition" of rational behavior. Sets of rationality postulates have been developed for different kinds of

[6] Hedström (2005) has introduced "DBO theory" of action, with "DBO" for "desires, beliefs, and opportunities." DBO theory, in turn, is the theoretical core of "analytical sociology" à la Hedström. It is tempting to conceive of DBO theory as a useful umbrella term for variants of micro-theories that provide alternative and possibly competing elaborations of the idea of goal-oriented and incentive-driven behavior. It is likewise tempting to use Hedström's label "analytical sociology" rather than "rigorous sociology" as an umbrella term for the family of research programs considered here. However, rather than seeing learning theories and theories of rational choice as *variants* of DBO theory, Hedström himself (see, for example, 2005: 41) prefers to see DBO theory as an *alternative* to micro-theories such as learning theories or theories of rational choice. Similarly, he promotes analytical sociology as an *alternative* to other variants of rigorous sociology. In my view, Hedström's positioning of DBO theory and analytical sociology is unfortunate because it increases the heterogeneity of the discipline rather than help spelling out and expanding its common core.

situations, including individual decision theory, which characterizes rational choice of an "isolated individual" in "parametric situations" (choice under certainty, risk, or uncertainty), and game theory, which characterizes rational behavior in a social setting with actors that are interdependent in the sense that the outcomes of an actor's behavior also depend on the behavior of other actors and vice versa ("strategic situations").

From the perspective of systematic theory formation, it is an important but unfortunately still often overlooked point – certainly so in discussions of the rational choice approach in sociology – that rational choice theories do *not* by any means start from an assumption of utility maximization. On the contrary, rational choice theories start from the assumption that observable behavior fulfills the rationality postulates, with utility maximizing behavior as an *implication* of these postulates. This likewise shows that one need not assume that actors consciously calculate (expected) utilities for their behavioral alternatives. In a rather strict sense, the rationality postulates imply nothing more than that actors who behave according to the theory are behaving *as if* they are maximizing a utility function.

Another point that is still often overlooked is that rational choice theories, while assuming certain formal properties of preferences such as transitivity, are neutral with respect to further substantive assumptions about these preferences. By way of example, while assuming that an actor who prefers A over B and B over C, will also prefer A over C, rational choice theories as such do not presuppose that A is indeed preferred over B and B over C, nor do they comprise assumptions about specific properties of A, B, and C that cause A to be preferred over B and B over C. In particular, rational choice theories *as such* do not assume at all that actors are exclusively interested in certain material consequences for themselves and are purely self-regarding in the sense of "utility = own money." Of course, empirical applications of these theories require additional substantive assumptions, alongside the rationality postulates or, respectively, the assumption of utility maximization, about actors' preferences, as well as substantive assumptions about, for example, their behavioral alternatives and their beliefs concerning the consequences of choosing a certain alternative (see Raub 1984: Chapter 2.2.). However, these are assumptions that should not be confused with the primary and secondary definitions of rationality. It is true that rational choice assumptions are often combined with assumptions that actors are purely self-regarding, so much so that the combination of both types of assumptions is sometimes referred to as the "standard model." This should not distract from the fact that rationality assumptions on the one hand and assumptions about self-regarding preferences on the other are distinct assumptions of the standard model.



Gary Becker

Coleman as well as Gary Becker have championed the application of rational choice theory in sociology.[7] Indeed, rational choice sociology has become an influential, though controversial, paradigm of rigorous sociology and is closely associated with purposive action explanations in the spirit of Coleman's diagram. Why is it that rational choice theories are prominent in rigorous sociology and have become a standard for purposive action explanations, notwithstanding empirical regularities from a variety of contexts that are hard to reconcile with assumptions about rational behavior (see Kahneman and Tversky 2000 for an overview of the extensive work on "anomalies" in situations without interdependence between actors and Camerer 2003 for an overview of related work in strategic situations)? Part of the answer is that they are quite often rather successful at explaining micro- and macro-level outcomes, including macro-level regularities. This includes testable predictions as well as a reasonable amount of corroborating empirical evidence, certainly in comparison with competing theories (see, for example, Wittek, Snijders, and Nee 2013), also showing that Green and Shapiro's (1994) influential argument on the gap between rational choice theories and empirical research is not as well justified now as it was 20 years ago.

However, the prominence of rational choice approaches in rigorous sociology is likewise due to another feature. Rigorous sociology emphasizes the importance of analytical rigor in the sense of providing *arguments*: sociology is not only about theory on the one hand and on empirical evidence on the other. It is also – and importantly – about deriving implications, including testable implications, from assumptions. Thus, theory formation and explanation are also – and importantly – about the *link* between assumptions and empirical evidence. Generating implications is often far from trivial, certainly so when it comes to micro-macro links and to deriving micro- and macro-level outcomes, including macro-level regularities, from a complex set of assumptions at the macro- as well as micro-level, including bridge assumptions and transformation rules on how both levels are related (see, for example, Coleman 1964a; Ziegler 1972; Hummell 1973). It

[7] Becker, an economist at the University of Chicago, held a joint appointment at the Chicago Department of Sociology. With Coleman, he ran the "Rational Models Seminar" for many years that engaged scholars from economics, sociology, psychology, philosophy, law – anything connected with the insights and limits of applying theories of rational choice to social, economic, and political issues. The seminar was a major attraction during my visiting appointments at UoC in the 1990s, including the experience of being "grilled" three times as presenter by Becker, Coleman, and others.

is a strength of rational choice theories that they provide parsimonious assumptions on regularities of human behavior that lend themselves to micro-macro models because they keep such models tractable and allow for the derivation of implications from micro-macro models, certainly when rationality assumptions are combined with assumptions about self-regarding preferences in the standard model. It seems a welcome side-effect that the use of similar rational choice theories in various micro-macro models that address quite different phenomena at the macro-level allows for a common theoretical core and coherence of such models as well as the cumulative growth of knowledge (see Diekmann and Voss 2004: 20). Rational choice theories are a tool for preserving sufficient family resemblance over a series of middle-range theories.

One can argue (see Harsanyi 1976: Chapters 6 and 7; 1977: Chapters 1 and 2) that although rational choice models and more specifically the standard model sometimes yield empirically problematic conclusions, it is useful to employ them for establishing a "benchmark" so that empirically observed refutations become themselves explananda for more refined models. Thus conceived, rational choice models contribute to theoretical pluralism by providing interesting new explananda and touchstones for competing theories. Another argument (e.g., Goldthorpe 2000) is that errors in predicting individual behavior with a rational choice micro-model will cancel out on the macro-level. Finally, Becker (1976: Chapter 8) has developed the argument that typical macro-outcomes of micro-macro models are robust relative to replacing the rational choice micro-model by alternative micro-models. Coleman (1987b; 1990: 19) argues in a similar direction and suggests that replacing simple rational choice assumptions by more complex micro-theories would undermine the tractability of the micro-macro model in the sense that it becomes unfeasible to derive implications for micro- and, in particular, macro-outcomes at all, certainly so when bridge assumptions and transformation rules are involved in the model.

Such arguments are meanwhile far from undisputed. For example, theoretical as well as experimental work has shown that the "errors cancel out" argument is problematic and that macro-level implications need not be robust relative to variations in micro-level assumptions, certainly so in strategic situations with interdependent actors (see Camerer 2003 for an overview and Raub and Snijders 1997 for an example). It is therefore not surprising that alternative micro-theories have been developed and are being used in micro-macro models. Many of these alternatives preserve the idea of incentive-guided and goal-directed behavior. One can then distinguish two kinds of alternatives to standard rational choice micro-models. First, one can retain rational choice assumptions but replace standard additional assumptions such as the assumption of self-regarding behavior. Influential examples are models that assume inequity aversion or similar motives as elements of an actor's utility

functions as well as heterogeneity between actors with respect to the strength of such motives (see Fehr and Schmidt 2006; Fehr and Gintis 2007 for systematic surveys).

Second, there are micro-models that replace the rationality assumptions themselves and use alternative decision rules and behavioral heuristics, often modeling "bounded rationality." Such micro-models include, for example, a variety of models assuming myopic behavior, backward-looking learning models, and evolutionary models (see Macy and Flache 2009 for an overview and further references). Note that such micro-level assumptions are sometimes integrated in micro-macro models that allow for analytical solutions. Often however, these micro-level assumptions are incorporated into models that rely on computational simulation models.

It is still an open question whether, in what respects, or under what conditions such alternative micro-models should be seen as better alternatives to standard rational choice models and how broadly they can be applied. Also, note that the introduction of alternative micro-theories in micro-macro models leaves the question open of how much priority should be given to empirical tests and to empirical corroboration of micro-assumptions when the focus is on macro-outcomes. At the very least, it seems reasonable to consider modifying micro-foundations when actors' behavior deviates systematically from the behavior implied by the micro-theory and when deviations from the implications of the micro-theory have strong effects on macro-outcomes.

A remark on "quantitative" versus "qualitative" sociology

While much of empirical research in rigorous sociology employs quantitative methods, there is no systematic reason to assume that rigorous sociology is incompatible with more "qualitative" approaches, assuming for the moment that a distinction between "quantitative" and "qualitative" sociology makes sense. For example, Gambetta's study of the Sicilian Mafia (1993) and his study of violent Islamist extremists (Gambetta and Hertog 2016) are quite in line with core features of rigorous sociology and micro-macro models and likewise have a "qualitative" flavor. Employing research strategies similar to "implication analysis" (Lieberson and Horwich 2008), Gambetta shows that a rigorous "qualitative" approach is feasible by deriving implications for different features of the phenomena to be explained from the same set of assumptions, and by confronting these implications, as well as implications of competing (sets of) assumptions, with available empirical evidence concerning these different features.

Rational models for running an academic institution

A key feature of rigorous sociology is the focus on interdependence between actors that can lead to unintended consequences. Moreover, Popper has argued that knowledge develops via a trial-and-error process, by learning from mistakes, with the growth of knowledge being unforeseeable. These are core arguments for "piecemeal social engineering" (Popper 1945, 1957): adjusting institutions incrementally rather than radically redesigning them as a whole. Adjusting institutions step by step, comparing *de facto* results of such adjustments with expected results, facilitates learning from mistakes and making improvements, since undesired consequences can be more easily traced to their causes.

It seems reasonable to apply such principles in academic institutions, too.[8] In addition, academic institutions target the growth of scientific knowledge and thus depend on the creativity and intrinsic motivation of scholars and scientists as well as students. Creativity and intrinsic motivation, though, cannot be enforced or extorted. Piecemeal social engineering and the core features of academic institutions therefore suggest modesty in administrators in academia: accepting that opportunities to induce desired results directly through policy making are limited in an absolute sense as well as compared to the risks of producing unintended and undesired results; reluctance with respect to large-scale pre-designed plans; focusing instead on creating and maintaining an environment that facilitates the spontaneous emergence of new ideas and initiatives, and on identifying and providing support for such ideas and initiatives. One way of doing so is to provide support, including and especially material support such as extra funding, primarily when some progress has already been achieved spontaneously, rather than assuming that no support is needed in such cases and that resources can be better spent where progress would be desirable but has not been achieved.[9]

Given that theoretical pluralism and competition between theoretical ideas are favorable for the growth of knowledge and given that growth of knowledge depends on creativity and intrinsic motivation, careful selection on all levels – students, including PhD students, postdocs, junior and senior faculty – becomes a major requirement for establishing and maintaining a favorable environment for

[8] On the rules of thumb for running a Department outlined in this text box, I learned much from Reinhard Wippler, my predecessor as Utrecht Chair of Theoretical Sociology and leader of the Utrecht Department of Sociology (see his own farewell lecture Wippler 1996). Note, too, that the rather unassuming suggestions outlined here may usefully complement much more fundamental discussions of the future of the academic system such as Van der Zwaan (2017).

[9] Of course, taking into account that the Matthew effect ("success breeds success") in science can have unintended and undesired consequences (Merton 1968; see also Van de Rijt et al. 2014).

the growth of knowledge. Moreover, groups constituting the core environment for research, teaching, and education as well as groups members identify with are Departments (in the sense of an *afdeling* at Dutch universities), research groups, and groups with members jointly responsible for a set of teaching tasks. Such groups can flourish in the sense of being fertile ground for new ideas by combining a common core with diversity. A common core, for example in the sense of shared methodological standards such as those of rigorous sociology, facilitates intellectual exchange and criticism. A common core also mitigates fragmentation, that is, a lack of shared standards for assessing the relative merits of competing ideas.[10] Diversity then ensures that alternative and competing ideas are explored and that group members can learn from each other: this is diversity in the sense of complementary expertise of the members, including with respect to disciplinary background, as well as a variety of research fields and domains in which research is executed. Academic leaders of such groups clearly have an important task and responsibility in striking a sound balance with respect to such "unity in diversity."

Note that careful and open selection of new group members, with a focus on selection from outside rather than on internal promotions in the case of vacancies, will contribute to diversity. New group members from outside bring their own expertise and research fields and they also contribute to diversity through their network of contacts outside the group. It should be clear, too, that "internationalization" – attracting more students and faculty from abroad and stimulating more international experience of Dutch students and faculty – likewise contributes to diversity and, via diversity, fosters an intellectual and academic climate that is favorable for the growth of knowledge (Raub 2016).

Finally, the core ideas underlying rigorous sociology have obvious but nevertheless noteworthy implications for teaching programs in sociology. Namely, these programs should include a strong focus on analytical thinking. More precisely, students should become capable of assessing whether an argument is valid in the sense that implications do follow from assumptions and of coming up themselves with valid arguments. Basically, thus, students should learn what a proof is. A radical proposal would be to make a basic course in logic part of the academic core of every undergraduate sociology program and to ensure that students acquire at least a basic level of mathematical sophistication.

---

[10] Focusing on the situation in sociology, Goldthorpe (2016: 128) argues that it is an advantage of the situation in Europe compared to the U.S. that European Departments, comparatively, have more of a common core *within*, while "differences in conception of sociology and in its actual practice" are to be found primarily *between* Departments. In the U.S., these differences are seemingly also found within Departments, thus contributing to more fragmentation.

2. Rational choice models of trust and cooperation in social dilemmas[11]

Let us now have a closer look at rational models in the technical sense of using assumptions about rational behavior in sociology. I consider rational choice models of trust and cooperation in social dilemmas, including a sketch of findings from empirical research.

2.1 Theory

Trust problems

We have already encountered an example that includes properties of social dilemmas: the co-authors and their joint paper. I simplify the example in some respects to specify what I mean by "trust and cooperation in social dilemmas."

Let us assume that there are only two co-authors, Adam and Eve. Adam is a senior scholar and Eve's supervisor. Due to other obligations, Adam has only limited resources in terms of terms of time and effort available for the joint project. Eve can invest more time and effort and she is also decisive for the joint paper and its quality due to her expertise. If she invests sufficient time and effort, together with Adam's input in terms of his experience and overview of the field, the outcome would be a high-quality co-authored paper. If Eve does not invest time and effort, the joint paper will be a failure but Eve can complete another single-authored paper of medium quality. In this case, Adam would end up with no new paper on his list of publications. Given his position as supervisor and senior scholar, though Adam can determine the order of authorship for the joint paper. Adam prefers to be the first author, while being second author would still leave him better off than with no paper at all. Eve would most prefer to be first author of the high-quality joint paper, while preferring her single-authored paper to being only second author of the joint paper.

Our example represents a trust problem between the co-authors. I follow Coleman's (1990: 96–99) general characterization of these problems. They involve two actors, a trustor and a trustee. In our example, Eve is the trustor and Adam is the trustee. Trustor Eve must decide whether or not she[12] will place trust in Adam by working hard on the joint paper. Now consider the following features of trust

[11] Parts of this chapter draw on materials from Buskens and Raub (2013), Raub, Buskens, and Corten (2014), and Buskens, Frey, and Raub (2017).

[12] For readability, "she" refers to the trustor and "he" to the trustee in the discussion of trust problems.

problems. First, placing trust implies that the trustee can subsequently honor or abuse trust. In our example, Adam can honor trust by agreeing that Eve will be the first author. Second, if the trustee honors trust, the trustor is better off than if trust were not placed: Eve prefers being first author of the high-quality joint paper to the single-authored paper. Conversely, trustee Adam can abuse trust by claiming first authorship for himself. If trust is abused, the trustor is worse off than had trust not been placed: Eve would prefer the single-authored paper over being only second author of the joint paper. Third, through placing trust, the trustor transfers resources to the trustee without any "real commitment" (Coleman 1990: 98) by the trustee to honor trust. Fourth, there is a time lag between the actions of trustor and trustee. The trustor first decides whether or not to place 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 will honor trust. A fifth feature, less explicitly addressed by Coleman, is that the trustee may have an incentive to abuse trust since he prefers first to second authorship for the joint paper.[13] A sixth feature is that trustor and trustee are both better off if trust is placed and honored than if the trustor does not place trust – in this respect, trust is a "lubricant of a social system" (Arrow 1974: 24).

Trust problems are a feature of a considerable number of interactions in social and economic life. Consider helping behavior as an example of social exchange (Blau 1964). Providing help requires time and effort but is valuable for the receiver. Eve helps Adam today, assuming that Adam will help Eve tomorrow. If Adam indeed provides help tomorrow, both Eve and Adam are better off than if they do not help each other. However, Adam may be tempted to benefit from Eve's help today without providing help himself tomorrow. Eve may anticipate this and not provide help in the first place, leaving both Eve and Adam worse off than if they had helped each other.

In economic exchange, buyer Eve may be insufficiently informed on the quality of a good offered by seller Adam. If she 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 into the transaction, again leaving buyer and seller worse off than after a "smooth" transaction.

Consider an experimental setup for trust problems. The experiment involves Adam and Eve as subjects and they play a game. Depending on how they play the game, Adam and Eve gain points that are converted into money at a fixed exchange

[13] Trust can be problematic, too, because the trustor is uncertain about the abilities and competencies of the trustee. In the literature (e.g., Barber 1983), this is sometimes labeled as a problem of "confidence." Here, I focus on trust problems due to the trustee's *incentives* to abuse trust rather than his *abilities* to honor trust.

rate at the end of the experiment. Eve is endowed with 20 points that she can transfer to Adam or keep for herself. If she keeps the endowment for herself, the game ends. If Eve transfers, the experimenter triples the endowment and Adam receives 60 points. Adam then chooses between splitting the 60 points by returning 30 points to Eve or keeping 60 points for himself. Afterwards, the game ends. Note that if Adam splits, both Adam and Eve are better off than if Eve had decided not to transfer her endowment. However, keeping 60 points for himself ensures an even higher payoff from the experiment for Adam than splitting them, while Eve is then worse off than if she had decided not to transfer her endowment.

The extensive literature on trust focuses on two different issues (see Craswell 1993). First, in the spirit of Arrow's remark on trust as a lubricant of a social system, research addresses the *consequences of trust*, be it consequences on the level of individual behavior or the consequences for macro-level outcomes such as societal cohesion or economic prosperity. Assumptions about trust are then used as (part of) the explanation for 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, rather, undermine trust.

Game-theoretic analysis

Trustor and trustee are interdependent in trust problems. What happens to the trustor, assuming she places trust, depends on the behavior of the trustee. Conversely, the trustee's outcome depends on trustor behavior. Game theory provides tools for the analysis of such situations (Raub and Buskens 2006 provides a brief introduction to game theory, Rasmusen 2007 is a textbook accessible to readers with modest training in formal theoretical model building and no prior exposure to game theory). A simple game-theoretic model capturing the core features of trust problems as well as our examples is the standard *Trust Game* depicted in Figure 2 (Dasgupta 1988; Kreps 1990a; a tree-like representation like the one in Figure 2 is known as the "extensive form" of a game). To facilitate interpretation, Figure 2 also includes the numerical example representing the experimental set-up. The game involves a trustor (indexed "1") and a trustee (indexed "2"). The trustor moves first and must 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 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 2, R_i is equal to 30, S_1 equal to 0, and T_2 equal to 60 points. I also refer to placing trust as *trustfulness* and to honoring trust as *trustworthiness*, with "trust" sometimes referring to trustful and trustworthy behavior in trust problems.

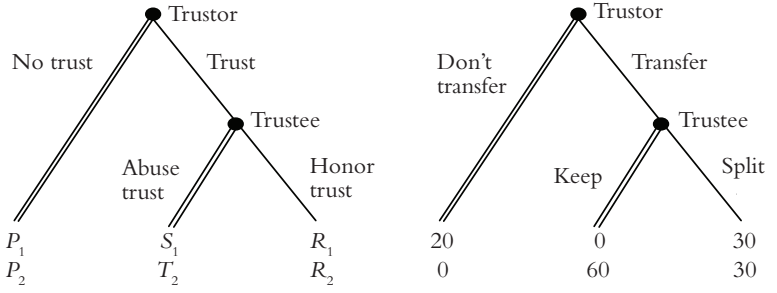


Figure 2: The Trust Game and an example.

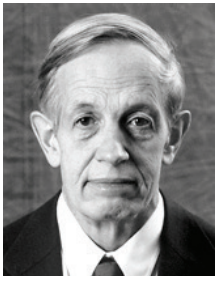
A standard game-theoretic assumption is that the payoffs actors receive at the end of the game represent *utilities*. Another is that all actors know the structure of the game (they know Figure 2), know that all actors know the structure of the game and so forth ("common knowledge"). The game is played *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).[14] We assume that the actors are rational in the sense that they maximize utility, given their expectations regarding the behavior of other actors, and actors assume that other actors are rational as well. Under these assumptions, 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.

It makes sense to have a somewhat closer look at game theory as a tool for analyzing behavior in strategic situations. Game theory assumes that actors behave as if they try to realize their preferences, taking their interdependencies as well as rational behavior of the other actors into account (e.g., Harsanyi 1976: Chapter 6). Thus, game theory is a tailor-made tool of rigorous sociology, since interdepend-

[14] Throughout, and for reasons that will become transparent, I focus on noncooperative games.

encies between actors and actors taking their interdependencies into account are likewise the core of Weber's (1947: 88, emphasis added) famous definition of social action: "Sociology […] is a science which attempts the interpretive understanding of social action in order thereby to arrive at a causal explanation of its course and effects […] Action is social in so far as […] *it takes account of the behaviour of others and is thereby oriented in its course*."

A core concept of game theory is an actor's *strategy*. Intuitively, this is a rule specifying the actor's behavior for each situation that can arise during a game and in which the actor has to take a decision. In the Trust Game, there is only one such decision for each actor. In general, a strategy is more "complex," namely, a plan describing how to behave during the game that specifies the actor's behavior for all circumstances that could emerge. The aim of game theory is to specify the *solution* of a game, that is, the strategies of rational actors in such a game. This is non-trivial in situations with interdependence between actors. After all, the consequences for an actor of choosing a certain strategy also depend on the strategies chosen by other actors and vice versa. Thus, an actor has to anticipate other actors' behavior and vice versa. The problem is that actors have to form rational expectations with respect to the behavior of other actors, whose behavior likewise depends on such expectations.



John Nash

Nash's (1951) equilibrium concept makes a key contribution towards tackling the problem. A *Nash equilibrium* is a combination of strategies, one for each actor, such that each actor's strategy maximizes the actor's expected payoff against the strategies of the other actors. Thus, no actor has an incentive to deviate unilaterally from his own equilibrium strategy, given that the other actors use their equilibrium strategies. One easily verifies that not placing trust, while trust would be abused, is an equilibrium of the Trust Game. In Figure 2, double lines indicate the actors' moves in the equilibrium. Simple, albeit strong, assumptions about properties of the solution of a noncooperative game imply that the solution has to be a Nash equilibrium. More specifically, assume (1) that a game has a unique solution (assume, thus, that the concept of rational behavior is well defined), (2) that all actors behave as if they anticipate the solution, and (3) that all actors are rational. It then follows that the solution has to be a Nash equilibrium. Therefore, Nash equilibrium behavior is the basic game-theoretic specification of rational behavior in strategic situations.

One of Nash's major achievements, the one that secured him the 1994 Nobel Prize in economics – together with Selten and Harsanyi – for his work in game theory is the proof of the existence of a Nash equilibrium for a broad class of games,

namely, finite games. These are games with a finite number of actors and a finite number of (pure) strategies for each actor. However, many games have more than one Nash equilibrium. Hence, being a Nash equilibrium is a necessary but not a sufficient condition for the solution of a game. This *equilibrium selection problem* is commonly seen as one, if not *the*, fundamental problem of game theory (see Kreps 1990b: Chapter 5 for an accessible discussion).

The equilibrium selection problem emerges in the Trust Game as soon as we allow for mixed strategies in the sense that actors can randomize when they must make a decision. One easily sees that the set of Nash equilibria is the set of all strategy combinations such that the trustor does not place trust while the trustee would honor trust with probability $0 < p \leq (P_1 - S_1)/(R_1 - S_1)$.[15]

Reinhard Selten

The equilibrium selection problem indicates that the solution of a game must fulfill additional criteria beyond being a Nash equilibrium. To this end, various *refinements* of the Nash equilibrium concept have been developed. The most important refinement is Selten's (1965) *subgame-perfect equilibrium*. For an intuitive understanding of subgame perfection, consider a part of a game tree like the one in Figure 2 that can itself be considered a game tree and, thus, represents a subgame of the original game. For example, in the Trust Game, the tree that starts with the node where the trustee chooses between honoring and abusing trust represents a subgame of the Trust Game. A subgame-perfect equilibrium is a strategy combination that is a Nash equilibrium for the game and *also* for each subgame. Selten proved the existence of subgame-perfect equilibria for finite games.

The intuitively attractive property of subgame-perfect equilibria is that they comprise credible promises as well as credible threats. Obviously, in the Trust Game, the equilibrium such that the trustee would abuse trust while the trustor does not place trust is the unique subgame-perfect equilibrium. Note that all equilibria of the Trust Game that are not subgame perfect comprise an (implicit) promise of the trustee that is not credible in the sense that a rational, that is, utility maximizing trustee would not be willing to honor his promise. More specifically, in a Nash equilibrium of the Trust Game that is not subgame perfect, the strategy of the trustee implies that he honors trust with positive probability, while this would not maximize his payoff at

[15] Although multiple equilibria exist in the Trust Game, they only differ in the specification of behavior at the node that will not be reached in equilibrium: behavior is the same in all equilibria of the Trust Game, namely, no trust is placed. In other games, though, different equilibria generally do differ with respect to the behavior of the actors that is induced by the respective equilibrium strategies.

that node. Subgame-perfect equilibria guarantee that promises and threats that are not credible never have to be executed. The intuitive justification of subgame perfection in the sense of credible threats and promises reveals that whether or not a rational actor uses a certain strategy may largely depend on what action the strategy requires in situations that will never actually emerge, given the solution. From a game-theoretic perspective, subgame-perfect equilibrium behavior is a natural further specification of rational behavior. Observe that double lines in Figure 2 indicate behavior induced by the subgame-perfect equilibrium. A double line that runs from the starting node of a game to an endnode indicates the equilibrium path of play, i.e., the behavior that results when the actors implement their subgame-perfect equilibrium strategies. The Trust Game has many Nash equilibria and a unique subgame-perfect equilibrium. In such a case, the subgame-perfect equilibrium can be considered the solution of the game. In the following, we employ this equilibrium concept. However, we will see that games can have more than one – and possibly many – subgame-perfect equilibria. In general, therefore, subgame perfection is, at best, a necessary but not a sufficient property for the solution of a game.[16]

Defining social dilemmas

If trust is not placed in the Trust Game, both trustor and trustee are worse off than when trust is placed and honored. Technically speaking, the no-trust outcome is Pareto-suboptimal. As Rapoport (1974) put it, individual rationality in terms of equilibrium behavior leads to "collective irrationality" in the sense of Pareto-suboptimality. Such a "conflict" between individual and collective rationality is the core feature of a social dilemma and trust problems are a paradigmatic example of a social dilemma involving two actors. While "social dilemma" is a label commonly used in sociology and social psychology, such a situation is often referred to as a "problem of collective action" or the "tragedy of the commons" in political science and as a "public goods problem" in economics (see Ledyard 1995: 122).

Actors *cooperate* when they behave in a manner that produces a better outcome for each of them than in a situation where all or at least some actors abstain from cooperation. In the Trust Game, the trustor cooperates by placing trust and the trustee by honoring trust. A *social dilemma* is a situation with strategically interdependent actors in which at least some actors have individual opportunities and incentives to abstain from cooperation (to *defect*) and all actors are worse off when actors succumb to those incentives, compared to the cooperative outcome. In the Trust

[16] The well-known "second-order free-rider problem" (e.g., Coleman 1990: Chapter 11) refers to the problem of providing sufficient incentives for the implementation of sanctions. Subgame-perfection of an equilibrium assures the solution of this problem for rational actors.

Game, the trustee defects when he abuses trust and not placing trust is defection by the trustor. More technically, when actors defect, the outcome is Pareto-suboptimal (there is another strategy combination in which each actor is better off), while cooperation is Pareto-optimal (there is no other strategy combination in which each actor is better off), and a Pareto-improvement compared to the outcome when actors defect (each actor is worse off when each actor defects than when everybody cooperates). Thus, a social dilemma game is a noncooperative game with a Pareto-suboptimal solution. Cooperation indicates a strategy combination that is associated with an outcome of the game that is Pareto-optimal and a Pareto-improvement compared to the solution, while cooperation is typically not an equilibrium outcome so that there are indeed actors with an incentive to deviate.[17]

Some game-theoretic models of social dilemmas

By now, trust problems have become a standard example of social dilemmas, with the Trust Game as an accompanying game-theoretic model. Many other social situations with strategically interdependent actors have dilemma properties, too, and a range of game-theoretic models representing such situations is available, including models that can account for trust problems with more complex features than considered so far.

For example, it is conceivable that a trust problem is more complex in the sense that trustor and trustee are not obliged to make binary choices (placing versus not placing trust, honoring versus abusing trust) but have a larger set of feasible actions. Durkheim ([1893] 1973: Book I, Chapter 7) discussed a trust problem in economic exchange in his analysis of the division of labor in society, albeit using other terminology. He highlighted the limits of the contractual governance of economic transactions. Governing a transaction exclusively via a contract would require that the present and future rights and obligations of the transaction partner are specified explicitly for all circumstances that might arise during and after the transaction. However, to design a contract covering all these contingencies is costly, thus reducing the gains from trade, or is not even feasible. Thus, mutually beneficial economic exchange – cooperation – presupposes the solution of a social dilemma arising from the problem of incomplete and implicit contracts (see Weber [1921] 1976: 409 for similar arguments in his sociology of law and Macaulay 1963 as a "modern classic"). A core feature of this dilemma is that the actors involved do not

[17] This conceptualization has been suggested by Raub and Voss (1986), derives from Harsanyi's (1977) now classic treatment, and is closely related to various other approaches (see Van de Rijt and Macy 2009 for an overview and discussion). Note that it is *not* so that noncooperative games exclude cooperation by definition. As we will see, much of the analysis of noncooperative game specifies conditions for cooperation in a noncooperative game.

make binary choices but have a larger set of feasible actions. When contemplating exchange, buyer and seller do not choose between "buying" and "not buying" and, respectively, selling a "peach" or a "lemon." Rather, the buyer now chooses how much time and effort to allocate to writing an externally enforceable contract that reduces the seller's opportunities for exploiting the buyer but likewise reduces the gains from trade. Conversely, the seller chooses the degree to which he behaves opportunistically by not sharing these gains. If the buyer anticipates "much" opportunism of the seller, he may prefer an extensive but costly contract that reduces the seller's opportunities for exploiting the buyer. Both actors, however, would be better off without costly contracting and with larger and shared gains from trade.

The *Investment Game* (Berg, Dickhaut, and McCabe 1995; Ortmann, Fitzgerald, and Boeing 2000) is a simple model for such situations. The trustor can now choose the degree to which she trusts the trustee and the trustee can choose the degree to which he honors trust. A variant of our experimental setup representing trust problems serves to illustrate. The trustor can transfer any number of points between 0 and 20 to the trustee. Transferring more points would mean that the buyer requires less extensive and thus less costly contractual safeguards. The buyer's "investment" is once again tripled, representing the gains from trade. Subsequently, the trustee decides which amount he will return to the trustor, namely, between 0 and the trustor's tripled investment. The seller thus decides how to share the gains from trade. The trustor's transfer indicates roughly how much she trusts the trustee, while sharing indicates the trustee's trustworthiness. It is obvious that the Investment Game has a unique subgame-perfect equilibrium: the trustee would never return anything, while the trustor would transfer nothing. At the same time, this outcome of the game is Pareto-suboptimal and both actors would be better off if, for example, the trustor had transferred all 20 points, with the trustee splitting the resulting 60 points. One sees that the game is a social dilemma game since rational actors forego all gains from trade.[18]

Incentive problems in exchange are often two-sided. For example, the seller has an incentive to sell a bad product for the price of a good one, while simultaneously the buyer has an incentive to delay payment. The *Prisoner's Dilemma*, presumably the most famous formal model for a social dilemma, represents two-sided incentive problems in exchange with binary choices for buyer and seller (Hardin 1982). In the Prisoner's Dilemma, depicted in Figure 3, actor 1 represents the row player, while actor 2 represents the column player. In this game, actors

[18] The substantive interpretation of certain behaviors in the Investment Game in terms of "trust" and "abuse of trust" is somewhat problematic. If the amount transferred by the trustor is "small," returning a "small" amount could also be interpreted as a punishment the trustee inflicts on the trustor for not trusting him rather than as abusing trust.

choose simultaneously[19] between cooperation and defection. Mutual defection is not only the unique equilibrium, defection is also a dominant strategy for each actor: whatever the strategy of the other actor, defection is always an actor's unique strategy that maximizes his payoffs. The Prisoner's Dilemma resembles the Trust Game and the Investment Game in that rational behavior of the actors involved implies a collectively irrational outcome. There is likewise an interesting difference, however. Both in the Trust Game and in the Investment Game, only one actor, the trustee, has incentives for opportunistic behavior, i.e., behavior that impairs the partner by exploiting the partner's cooperation. The trustor can foresee this and can decide not to place trust and, respectively, not to transfer anything. Thus, defection by the trustee is motivated by greed, while defection by the trustor is defensive and motivated by fear. In this sense, the Trust Game and the Investment Game are examples of social dilemmas with one-sided incentive problems. Unlike in the Trust Game and the Investment Game, the Prisoner's Dilemma represents two-sided incentive problems: *both* actors have an incentive to exploit the cooperation of the other actor since $T_i > R_i$ for each of them. Thus, *both* actors are motivated by greed *and* fear.

		Actor 2	
		Cooperation	Defection
Actor 1	Cooperation	R_1, R_2	S_1, T_2
	Defection	T_1, S_2	**P_1, P_2**

Figure 3: The Prisoner's Dilemma ($S_i < P_i < R_i < T_i$) the bold-faced cell indicates the unique equilibrium.

Social dilemmas can involve more than two actors. Examples include environmental public good problems (e.g., Hardin 1968) and Olson's (1965) case of the provision of public goods by organizations such as trade unions. We briefly discuss three examples of game-theoretic models for such dilemma situations. Taylor ([1976] 1987) has introduced an *n-actor Prisoner's Dilemma* that generalizes the two-actor version (see, e.g., Schelling 1978 for a more restrictive definition of the game) and, like the standard Prisoner's Dilemma, is a game with binary choice for each actor. Each actor can choose between cooperation and defection. Each

[19] "Simultaneous" in the sense that an actor, when making a choice, has no information on the choice of the other actor and vice versa: no actor can make his own behavior conditional on the other actor's behavior. It is thus not necessary to assume that the actors decide *at the same time*.

actor's payoff depends exclusively on his own choice and the number v ($0 \leq v \leq n - 1$) of other actors who cooperate. Since the game is a Prisoner's Dilemma, it is reasonable to assume that defection is a dominant strategy: independent of v, defection is the unique strategy that maximizes an actor's payoff. Hence, each actor has an incentive to defect, regardless of the behavior of others. A Prisoner's Dilemma likewise requires that each actor prefers cooperation of all actors to defection of all actors. In addition, Taylor assumes that each actor's payoff when he defects and at least one other actor cooperates is larger than his payoff when everybody defects. Under these assumptions, defection by each actor is the unique equilibrium of the game. Moreover, the equilibrium is Pareto-suboptimal. On the other hand, cooperation by each actor, while not an equilibrium, is a Pareto-improvement compared to universal defection and is also Pareto-optimal. Clearly, then, the n-actor Prisoner's Dilemma is a social dilemma game.

The *Public Goods Game* (e.g., Gächter and Thöni 2011) is a model for a social dilemma with n actors in which the actors do not have to make a binary choice between cooperation and defection. Rather, they can choose between more cooperation or less cooperation. Each actor i has an endowment E. Actors simultaneously choose a contribution g_i to the public good with $0 \leq g_i \leq E$. The total amount contributed is multiplied by m, with $1 < m < n$, and subsequently the result is divided equally among the actors. Hence, each actor's payoff depends on his own contribution and the contribution of all other actors. Since $m < n$, the individual return from i's own contribution is smaller than the individual contribution itself. Hence, contributing nothing ("free riding") is a dominant strategy for each actor and the game has a unique equilibrium, with nobody contributing anything. Obviously, this outcome is Pareto-suboptimal: since $m > 1$, each actor is better off and the outcome is Pareto-optimal when each actor contributes his whole endowment. This shows that the Public Goods Game is a social dilemma game with each actor having an incentive to defect.[20]

A final example of a social dilemma with n actors is Diekmann's (1985) *Volunteer's Dilemma*. Actors have binary choices. They decide simultaneously whether or not to provide a collective good. The good is costly and will be provided if at least one actor – the "volunteer" – decides to provide. Contributions by more than one actor are feasible and in that case each actor pays the full costs of providing the good but multiple-actor contributions do not affect the utility level of any actor. The Volunteer's Dilemma differs from the n-actor Prisoner's Dilemma and the Public

[20] Note that the two-actor Public Goods Game relates to the standard Prisoner's Dilemma with two actors much in the same way as the Investment Game relates to the Trust Game. The Investment Game is a variant of the Trust Game and the two-actor Public Goods Game is a variant of the Prisoner's Dilemma with continuous rather than binary choices for the actors.

Goods Game in that the costs of providing the collective good are smaller than the gains from the good. The matrix in Figure 4 summarizes the normal form of the game, with the rows representing an actor's strategies, namely, to provide the good (PROV) or not to provide (DON'T), with columns indicating the number of other actors who choose PROV, and cells representing the actor's payoffs as a function of his own strategy and the number of other actors who choose PROV.

	Number of other actors choosing PROV				
	0	1	2	…	$n-1$
PROV	$R-K$	$R-K$	$R-K$	…	$R-K$
DON'T	0	R	R	…	R

Figure 4: The Volunteer's Dilemma ($R > K > 0$; $n \geq 2$).

In the Volunteer's Dilemma, therefore, each actor has an incentive to provide the public good when nobody else is providing, while all other actors have an incentive not to provide if there is one volunteer. Diekmann (1985) identifies the bystander intervention and diffusion of responsibility problem (Darley and Latané 1968) as an example of a social situation for which the Volunteer's Dilemma is a reasonable model. This is a situation with n actors witnessing an accident or a crime. Everybody would feel relieved if at least one actor helped the victim, for example by calling the police. However, providing help is costly and each actor might be inclined to abstain from helping, hoping that someone else will help.

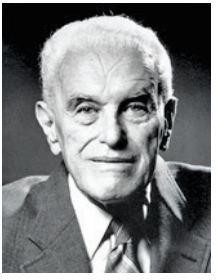
The Volunteer's Dilemma has n equilibria in pure strategies. These are the strategy combinations with exactly one volunteer choosing PROV while all other actors choose DON'T. Each of these equilibria is Pareto-optimal. However, the equilibria involve a bargaining problem, since each actor prefers the equilibria with another actor as the volunteer to the equilibrium where he himself is the volunteer. Moreover, while the game is symmetric, the n equilibria in pure strategies require that actors do not behave the same. It can be shown that the Volunteer's Dilemma has a unique symmetric equilibrium in mixed strategies such that each actor chooses DON'T with probability $p^* = (\frac{K}{R})^{\frac{1}{n-1}}$. It follows from elementary properties of equilibria in mixed strategies that each actor's expected payoff associated with the symmetric equilibrium in mixed strategies is $R - K$. Hence, each of the equilibria in pure strategies is a weak Pareto-improvement since in each of those equilibria the volunteer is not worse off while all other actors are better off. From a game-theoretic perspective, the symmetric equilibrium in mixed strategies is a plausible candidate for the solution of the Volunteer's Dilemma.

An important feature of the Volunteer's Dilemma is that there is indeed another combination of mixed strategies that is symmetric so that each actor chooses the same strategy, namely choosing DON'T with probability $p^{**} = (\frac{K}{nR})^{\frac{1}{n-1}}$. While not an equilibrium of the Volunteer's Dilemma, this combination of mixed strategies is Pareto-optimal and is associated with expected payoffs that are higher for each actor than his expected payoff in the symmetric mixed strategy equilibrium. Since the strategy combination whereby each actor chooses DON'T with probability p^{**} is not an equilibrium, it follows that the Volunteer's Dilemma is indeed a social dilemma game. The mixed strategy of choosing DON'T with probability p^{*} represents defection, while cooperation would imply that each actor chooses DON'T with probability p^{**}. Note that $p^{*} > p^{**}$, that is, cooperation in the Volunteer's Dilemma implies a larger individual probability of providing the public good than defection.[21]

Toward more complex game-theoretic models

Game theory provides flexible tools for analyzing social dilemmas not only because it makes different models available for different kinds of social situations, it also allows for the inclusion of additional features in the analysis. For example, game-theoretic models make it possible to account for *chance events* ("external contingencies") that can emerge in strategic situations and can affect actors' behavior. *Information issues* can likewise be included. The models considered so far assume that actors are completely informed about all features of the game. This assumption is often problematic for social interactions. For example, in trust problems, the trustor might be incompletely informed about the trustee's incentives. More specifically, the trustee might derive more utility from honoring trust than from abusing trust due to the internalization of norms and values that imply "internal sanctions" when trust is abused. Or, the trustor might be incompletely informed about the trustee's alternatives in the sense that the trustee might not have any opportunity to abuse trust. Under such *incomplete information*, the trustor can no longer be sure about the behavior of a rational trustee after placement of trust. Game theory allows for modeling incomplete information and analyzing rational behavior under incomplete information (see Rasmusen 2007: Chapter 2 for an introduction to the theory of games with incomplete information and Bacharach and Gambetta 2001 for arguments as to why the Trust Game with incomplete information may be a more adequate model of trust problems than the Trust Game with complete information in Figure 2).

[21] See Raub, Buskens, and Corten (2014) for a more detailed discussion of the games mentioned as well as other game-theoretic models for social dilemmas.



John Harsanyi

Consider a Trust Game where the trustor has incomplete information about the incentives of the trustee (incomplete information about the trustee's behavioral opportunities would imply similar results) and that includes a chance event. The game starts with a random move of Nature that determines the *type* of trustee a trustor encounters. With probability π, the trustee's payoff from abusing trust is $T_2^* < R_2$ (e.g., this trustee has internalized norms and values of trustworthiness), while with probability $1 - \pi$ the trustee's payoff from abusing trust is $T_2 > R_2$. The trustee knows his own incentives, while the trustor is only informed on the probability π and cannot directly observe the outcome of the random move of Nature. It is straightforward to identify the subgame-perfect equilibrium of the Trust Game with incomplete information. A trustee with internalized norms and values of trustworthiness honors trust, while the other type of trustee abuses trust. If the trustor does not place trust, she receives P_1, while her expected payoff from placing trust is $\pi R_1 + (1 - \pi)S_1$. The trustor's unique equilibrium strategy is therefore not to place trust if $\pi < (P_1 - S_1)/(R_1 - S_1)$. Conversely, placing trust is the unique equilibrium strategy if $\pi > (P_1 - S_1)/(R_1 - S_1)$. If $\pi = (P_1 - S_1)/(R_1 - S_1)$, the game has multiple subgame-perfect equilibria since placing and not placing trust yield the same expected payoff.[22] Note that $(P_1 - S_1)/(R_1 - S_1)$ is a convenient measure of the risk the trustor incurs when placing trust. Our example of a game with incomplete information is simple. Harsanyi (1967/68) has pioneered the study of games with incomplete information and has shown how they can be analyzed in general.[23]

In a social dilemma, cooperation is beneficial for the actors directly involved. It is important to note, however, that it need not be beneficial for third parties. For example, members of the Mafia (e.g., Gambetta 1993) or cartel members (e.g., Stigler 1964) are involved in social dilemma-like interactions, with cooperation being beneficial for the members and detrimental for victims and consumers. Thus from a societal perspective and from the perspective of policy making, the issue is not always how to induce and maintain cooperation. In cases such as the Mafia or cartels, undermining cooperation is desirable.

---

[22] Coleman's (1990: Chapter 5) well-known condition for placing trust is equivalent to $\pi > (P_1 - S_1)/(R_1 - S_1)$. It is thus not necessary to introduce this condition as an assumption; rather, it can be derived from a game-theoretic model as an implication.

[23] Analyses of games with incomplete information often involve particularly strong rationality assumptions (e.g., about Bayesian updating) that may be considered questionable as empirical assumptions about human behavior.

Parsons' challenge

Social dilemmas and cooperation are revered topics of sociology and are a strategic research site, in Merton's (1973) sense, for rigorous sociology and rational models. First, social dilemmas are a paradigmatic example of unintended consequences of goal-directed behavior: actors try to further their own interests and in doing so produce an outcome that is worse for all than an outcome they could have obtained by cooperating. Also, Pareto-optimality and, respectively, suboptimality are paradigmatic examples of macro-outcomes of behavior in interdependent settings in the sense that cooperation and defection are phenomena on the micro-level of individual behavior, while Pareto-optimality and -suboptimality are properties of the social system formed by the actors. Hence, theories and models of social dilemmas and cooperation exemplify the study of micro-macro links in rigorous sociology.

Second, social dilemmas are closely related to Hobbes' ([1651] 1991: Chapter 13) "naturall condition of mankind," with interdependent actors in a world of scarcity without binding and externally enforced contracts. In this world, actors can end up in the "warre of every man against every man." In a peaceful situation – in our terminology: a situation in which actors cooperate – everybody would be better off. This a social dilemma involving many actors (see Taylor [1976] 1987). Parsons (1937) considered the Hobbesian "problem of order" as "the most fundamental empirical difficulty of utilitarian thought" (1937: 91) and posed the challenge to specify conditions such that rational, i.e., incentive-guided and goal-directed actors can cooperate and thus avoid a life that is "solitary, poore, nasty, brutish, and short," as Hobbes vividly described it.

Approaches to social dilemmas and cooperation using game-theoretic tools take up Parsons' challenge and indeed try to specify conditions for the cooperation of rational actors in social dilemmas (Raub and Voss 1986). Such approaches avoid the "normative solution" proposed in various versions by Parsons and others, specifically in the functionalist tradition. According to the normative solution, cooperation is a result of internalized and shared norms and values as well as norm-conforming behavior. However, this approach only shifts the problem to the explanation of how such norms and values emerge and are maintained and is hardly compatible with the observation that the degree of norm-conformity varies not only between actors but also for the same actors over time. Game-theoretic approaches instead follow Coleman's (1964b: 166–167) radical suggestion for taking up Parsons' challenge: "Hobbes took as problematic what most contemporary sociologists take as given: that a society can exist at all, despite the fact that individuals are born into it wholly self-concerned, and in fact remain largely self-concerned throughout their existence.

Instead, sociologists have characteristically taken as their starting point a social system in which norms exist, and individuals are largely governed by those norms. Such a strategy views norms as the governors of social behavior, and thus neatly bypasses the difficult problem that Hobbes posed [...] I will proceed in precisely the opposite fashion [...] I will make an opposite error, but one which may prove more fruitful [...] I will start with an image of man as wholly free: unsocialized, entirely self-interested, not constrained by norms of a system, but only rationally calculating to further his own self interest."

Employing noncooperative games as models for social dilemmas takes up Parsons' challenge in yet another way. Binding agreements or binding unilateral commitments that are not explicitly modeled as the results of actors' decisions are excluded in noncooperative games. Technically speaking, such agreements and commitments must be modeled as moves in the extensive form of the game. Thus, noncooperative games model Hobbes' state of nature. In this way, one avoids shifting the problem to the explanation of the emergence and maintenance of external enforcement of agreements and commitments. In fact, the availability of external enforcement presupposes that a social dilemma has been solved because actors can benefit from such enforcement even if they have not contributed to the costs of providing enforcement. Assuming a noncooperative game implies that external enforcement, if it enters the analysis, must itself be explained in the first place.

Two types of research questions

In the following, I focus on two types of research questions on cooperation in social dilemmas. The first concerns conditions for cooperation in social dilemmas. What conditions make cooperation more likely? The second type of questions follows from the fact that cooperation, while not an equilibrium outcome of a social dilemma, is beneficial for the actors involved. Therefore, in principle, rational actors have incentives to try and change the strategic situation so that cooperation does become an outcome of equilibrium behavior. This leads to questions about how actors cope with social dilemmas: when and how do actors actively shape conditions for cooperation? In a sense, these are theoretically deeper questions about whether actors, when feasible, modify their strategic interdependence ex ante with an eye on the ex post effects of such modifications for subsequent cooperation (see Prendergast 1999 for a similar argument in a different but related context: the design of compensation contracts by employers to align the interests of employees). Both types of questions will be discussed for the special case of trust problems and the Trust Game. On the one hand, the focus on the Trust Game helps answer both types of questions due to its simplicity. On the other hand, many of the results for

the Trust Game generalize neatly to a broader class of games and thus to other social dilemma situations as well.[24]

Embeddedness effects on cooperation

Before, I distinguished two alternatives to standard rational choice models that combine rationality assumptions with assumptions on purely self-regarding preferences. One approach retains the rationality assumption and employs alternative assumptions about preferences. Another approach replaces the rationality assumptions. In terms of Coleman's diagram, a common feature of both approaches is that they focus on assumptions on the micro-level and on making such assumptions more complex in one way or the other. I now turn to research on trust that follows a different strategy and one that is, in a sense, more sociological in nature (see Wittek, Snijders, and Nee 2013: Introduction for a related discussion). This strategy aims to use more complex assumptions about the social context.[25] In terms of Coleman's diagram, this strategy aims to modify assumptions on the macro-level as well as bridge assumptions and transformationn rules that link the macro- and the micro-level. The paradigmatic application of the standard rational choice model in neoclassical economics assumed "atomized interactions" on perfect markets, that is, a context where strategic interdependence can be neglected.[26] In the following, macro-level assumptions that characterize perfect markets are replaced. I combine strong assumptions about individual rationality and individual preferences – in fact, the assumptions of the standard model – with assumptions about strategic interdependence between actors due to the "embeddedness" of action in ongoing relations and networks of relations. I thus show that embeddedness crucially affects the behavior of rational actors, including purely self-regarding actors, in social dilemmas.

Roughly, embeddedness (Granovetter 1985; Raub and Weesie 1992; Raub 1997) can mean that the actors involved in a Trust Game maintain an ongoing relationship with prior and expected future interactions. This is *dyadic embeddedness*. An example is that the co-authors working on a joint paper have already worked together in the past and furthermore expect that they might work together on other projects in

[24] The NWO PIONIER-program *The Management of Matches* (Raub and Weesie 1992; Raub 1997) which benefited from additional funding by Utrecht University and the Faculty of Social and Behavioral Sciences of Utrecht University provided very attractive conditions for research on both types of questions for many years.

[25] Ostrom (2003) sketches first steps towards a theoretical framework combining the different strategies.

[26] Of course, as every thorough textbook shows (e.g., Mas-Colell, Whinston, and Green 1995), neoclassical economics has much more to offer than only a theory of perfect markets.

the future. Moreover, a focal Trust Game can be related to interactions of trustor or trustee with third parties. This is *network embeddedness*. For example, Adam and Eve are members of the same Department and maintain contacts with other colleagues in their Department.

I distinguish two mechanisms, control and learning, through which dyadic and network embeddedness may affect trust (Raub 1997). *Control* refers to the case in which the trustee has short-term incentives for abusing trust, while long-term consequences of his behavior in the focal Trust Game depend on the behavior of the trustor. More precisely, if the trustee honors trust in the focal Trust Game, the trustor may be able to reward this by applying positive sanctions in the future. In our co-author example, Eve is willing to embark on a new joint project with Adam if Adam cooperates with respect to the order of authorship of the joint paper they are writing. Conversely, if the trustee abuses trust in the focal Trust Game, the trustor may be able to punish this by applying negative sanctions. Given dyadic embeddedness, the trustee must take into account that honoring trust in the focal Trust Game may affect whether or not the trustor places trust again in the future. Given network embeddedness, the trustee must consider that a trustor can inform third parties about the trustee's behavior in the focal Trust Game, for example other trustors with whom the trustee may be involved in future Trust Games. Again, whether or not other trustors are willing to trust the trustee may depend on honoring or abusing trust in the focal Trust Game. Thus, the trustee must trade off the short-term incentives to abuse trust against the long-term benefits of honoring trust and the long-term costs of abusing trust. This mechanism is known as conditional cooperation (Taylor [1976] 1987) or reciprocity (Gouldner 1960; Blau 1964; Diekmann 2004).

Embeddedness may affect trust through a second mechanism, namely, *learning*. I mentioned earlier that the trustor need not be completely informed about the trustee's behavioral alternatives and incentives. The trustor's beliefs concerning the trustee's characteristics can be affected by information on past interactions. This information can be obtained from past interactions of the trustor and trustee, that is, through dyadic embeddedness. Given network embeddedness, information can also be obtained from third parties who have interacted with the trustee in the past. If a trustee has been trustworthy in past interactions, a trustor might be more convinced that the trustee will be trustworthy again in the focal Trust Game than if information on the trustee's past untrustworthy behavior has been revealed. Table 1 summarizes the distinction between dyadic and network embeddedness as well as between learning and control (Buskens and Raub 2002; see Yamagishi and Yamagishi 1994: 138–139 for a similar discussion of learning and control effects through network embeddedness).

Table 1: Types of embeddedness and mechanisms through which embeddedness affects trust.

Two mechanisms	Two types of embeddedness	
	Dyad	Network
Control	Sanctioning possibilities of the trustor without involving third parties.	Sanctioning possibilities of the trustor that involve third parties.
Learning	Information about the trustee from past experiences of the trustor.	Information about the trustee from third parties.

Embeddedness effects are a common theme of the sociological literature. However, many authors neglect to clearly disentangle different types of embeddedness effects and the underlying mechanisms, either theoretically or empirically. Game-theoretic tools allow us to model embeddedness and, more importantly, to derive predictions concerning effects of embeddedness on trust.

This approach is in line with Coleman's (1987b) heuristic advice on developing and improving micro-macro models by combining robust assumptions on rational behavior with more complex assumptions on social structure. Granovetter (1985) also advocated precisely such a combination of assumptions in his influential programmatic pamphlet on the importance of accounting for embeddedness. Granovetter's criticism of the shortcomings of the neo-classical model of perfect markets with atomized actors has often been taken to imply that one had better abandon rational choice models in favor of more "realistic," socially inspired models of man. What has been widely overlooked, however, is that Granovetter sharply opposes "psychological revisionism," characterizing it as "an attempt to reform economic theory by abandoning an absolute assumption of rational decision making" (1985: 505). Rather, he suggests maintaining the rationality assumption: "[W]hile the assumption of rational action must always be problematic, it is a good working hypothesis that should not easily be abandoned. What looks to the analyst like nonrational behavior may be quite sensible when situational constraints, especially those of embeddedness, are fully appreciated" (1985: 506). He argues that investments in tracing the effects of embeddedness are more promising than investments in the modification of the rationality assumption: "My claim is that however naive that psychology [of rational choice] may be, this is not where the main difficulty lies – it is rather in the neglect of social structure" (1985: 506).

Reciprocity: repeated interactions and conditional cooperation

To see how trust can result from purely selfish rational actors who are "enlightened" in the sense that they take the long-term effects of their behavior into account, we consider a simple model of control effects through dyadic embeddedness, namely, Kreps' (1990a) model of a repeated Trust Game. In this model, the Trust Game is played repeatedly in rounds 1,2,…,*t*,… By way of example, our co-authors are repeatedly involved in joint projects that come with trust problems. More precisely, after each round *t*, another round $t + 1$ is played with probability w $(0 < w < 1)$, while the repeated game ends after each round with probability $1 - w$. The focal Trust Game is thus embedded in a more complex game in which the *Trust Game is repeated indefinitely*. In each round, trustor and trustee observe each other's behavior. In the repeated game, a strategy is a rule that prescribes an actor's behavior in each round *t* as a function of the behavior of both actors in the previous rounds (it is now obvious that a strategy for the repeated game is different from and, in an intuitive sense, more "complex" than a move in a round of the repeated game). An actor's expected payoff for the indefinitely repeated Trust Game is the discounted sum of the actor's payoffs in each round, with the continuation probability w as discount parameter. For example, a trustor who places trust throughout the repeated game, with trust being honored throughout, receives payoff $R_1 + wR_1 + \ldots + w^{t-1}R_1 + \ldots = R_1/(1 - w)$. Using Axelrod's (1984) label, the continuation probability w represents the "shadow of the future": the larger w is, the more an actor's payoff from the repeated game depends on what the actor receives in future rounds.

In the indefinitely repeated Trust Game, the trustor can exercise control. She can use a conditional strategy that rewards a trustee who honors trust in a focal Trust Game by placing trust again in future games. Such a conditional strategy therefore implies a promise. Conversely, the trustor can use a conditional strategy that punishes abuse of trust on the part of the trustee in the focal Trust Game by not placing trust in at least some future games. The conditional strategy thus also implies a threat.

If the trustor uses reciprocity in the sense of implementing a conditional strategy, the trustee can gain T_2 rather than R_2 in the current Trust Game by abusing trust. However, abusing trust will then be associated with obtaining only P_2 in (some) future encounters, while honoring trust will result in larger payoffs than P_2 in those encounters. Moreover, the larger the shadow of the future, the more important are the long-term effects of present behavior. Anticipating that the trustor may use a conditional strategy, the trustee must balance short-term $(T_2 - R_2)$ and long-term $(R_2 - P_2)$ incentives. It can be shown that reciprocity can be a basis for rational trust in the sense that the indefinitely repeated Trust Game has an equilibrium that is "cooperative" in that trust is placed and honored in each round. Consider

conditional behavior of the trustor that presents the trustee with the largest reward for trustworthy behavior and exacts the most severe punishment for untrustworthy behavior. This occurs if trust is placed 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 is any deviation from the pattern "place and honor trust") maximizes the trustee's payoff against such conditional behavior of the trustor if and only if

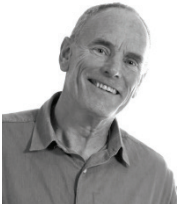
(1) $w \geq (T_2 - R_2)/(T_2 - P_2)$.

This condition requires the shadow of the future to be large enough compared to $(T_2 - R_2)/(T_2 - P_2)$, a convenient measure for a selfish trustee's temptation to abuse trust. A cooperative equilibrium can be likewise subgame-perfect. This implies that the trustor's promise to reward the trustee's trustworthy behavior by placing trust again in the future, and her threat to punish abuse of trust by not placing trust, are credible. Note that promises and threats can remain completely implicit: the trustor need not announce them verbally or otherwise. Moreover, given the credibility of the threat, a rational trustee has no incentive to deviate from the behavioral pattern implied by the equilibrium and hence the threat, being credible, need not be implemented. One sees that rational actors may cooperate in a noncooperative game. This shows that enlightened self-interest can indeed drive reciprocity and form a basis for trust among rational actors, in the sense that placing and honoring trust are due to equilibrium behavior. Reciprocity can be driven exclusively by the long-term, "enlightened" self-interest of the actors.

Placing trust conditionally and honoring trust in the sense of the strategies mentioned can be seen as following an informal norm (Voss 2001) 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 a norm or institution because each actor maximizes his or her expected payoff, 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). Moreover, since the norm prescribes conditional behavior, one sees that a core claim of functionalist sociology can be abandoned, namely, that cooperation must be independent of the partner's behavior by following an internalized norm of unconditional cooperation (Voss 1985). Coleman (1964b: 180) seemingly intuited this result when he argued that an important feature of socialization is "coming to

see the long-term consequences to oneself of particular strategies of action" rather than the internalization of norms.

The result for the indefinitely repeated Trust Game can be generalized. For example, an analogous result holds for an indefinitely repeated Investment Game. Friedman (1971, 1990) shows an analogous result for a broad class of indefinitely repeated 2- and *n*-person games. Roughly, if a social dilemma is repeated indefinitely and the shadow of the future is large enough relative to the incentives of the actors, that is, if a variant of condition (1) is fulfilled, then there is a cooperative equilibrium of the indefinitely repeated game, namely, an equilibrium that implies that the actors cooperate throughout the repeated game. The equilibrium of the repeated game induces a Pareto-optimal outcome and a Pareto-improvement compared to the Pareto-suboptimal solution of the original dilemma. Taylor ([1976] 1987) recognized the implications of these results for political philosophy and political science and used them for a rigorous analysis of Hobbesian arguments for the necessity of government.[27] From a political science perspective, Axelrod (1984) popularized the results in his influential study on the "evolution of cooperation." Voss (1982, 1985) seems to be the first sociologist who realized explicitly that the theory of repeated games has important implications for the problem of order and cooperation in social dilemmas.



Michael Taylor

Robert Axelrod

Cooperative equilibria, however, are not unique. For example, never placing trust and placed trust always being abused is always an equilibrium of the indefinitely repeated game. The "folk theorem" (e.g., Fudenberg and Maskin 1986) for repeated games implies that the indefinitely repeated Trust Game has many other equilibria, too, for large enough *w*. Thus, an equilibrium selection problem emerges. A typical, though sometimes implicit, argument in the literature on equilibrium selection in this context is "payoff dominance." An equilibrium is payoff dominated if there is another equilibrium that is associated with higher payoffs for at least some actors and is not associated with lower payoffs for any actor. In the indefinitely repeated Trust Game, an equilibrium that implies placed and honored trust throughout the game is evidently not payoff dominated by other equilibria, whereas the no-trust-throughout equilibrium is payoff dominated. More generally, the folk theorem highlights that

[27] Interestingly, Taylor wrote his book at NIAS, the *Netherlands Institute for Advanced Study in the Humanities and Social Sciences*, indicating how Dutch academic institutions not primarily aiming at "valorization" and short-term economic "impact" contribute to the growth of knowledge, including truly path-breaking work.

repeated games, in general, come with a coordination problem (see, e.g., Camerer 2003: Chapter 7) for rational actors, namely the problem that the actors involved must anticipate or learn which equilibrium to play.

For a balanced assessment of the merits of the model of conditional cooperation due to dyadic embeddedness it is useful to consider some shortcomings. For example, cooperative equilibria imply that trust is *always* placed and honored, whereas one might expect less than "perfect" trust, even under favorable conditions for trust. One can show (see, e.g., Taylor [1976] 1987) that there are equilibria that induce placement of trust and honoring trust only in some, rather than in all rounds of the game, with this pattern of behavior again being backed by a variant of conditional strategies. In that case, however, it is much more difficult to select one out of a wealth of such equilibria as a "solution candidate." Moreover, although no deviations occur in a cooperative equilibrium, many cooperative equilibria imply that trust *would* break down *completely* after the first deviation. This counter-intuitive feature can be circumvented, for example, by considering a game with imperfect monitoring (see Green and Porter 1984 for early work and Bernheim and Madsen 2017 for a recent contribution). Assume that the trustor, after placing trust, cannot observe the trustee's behavior, but can only observe the outcome of that behavior. This outcome, in turn, depends on the trustee's behavior but also on chance: a low payoff for the trustor after placement of trust can be due to abuse of trust by the trustee, but can also be due to "bad luck." The trustor must now solve an "optimal punishment" problem. If the trustor never punishes, or is too lenient in her punishments, a rational trustee would always abuse trust. But making the punishments too severe implies losses for trustor and trustee that go beyond what is necessary in terms of deterring the trustee from abusing trust. Equilibrium behavior that generates some honored trust throughout the game now requires the trustor to punish the trustee by withholding trust occasionally rather than constantly.

Testable predictions on embeddedness effects

How can we use game-theoretic models to derive testable predictions on embeddedness effects? Relating the game-theoretic analysis of the repeated Trust Game to Coleman's diagram offers a useful approach to answering this question and to intuiting how game theory can help generate testable prediction from micro-macro models in other contexts as well (Raub and Buskens 2006: 569–570; Buskens and Raub 2013: 119–120, 125). First, the specification of a game like the Trust Game 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 non-

cooperatively, assumptions about the sequence in which the actors move, about the shadow of the future (the probability w of a subsequent round of the Trust Game), and about condition (1) which reflects how the shadow of the future relates to the trustee's temptation to abuse trust. The game tree for the Trust Game also comprises assumptions about the actors' preferences over outcomes, represented by their payoffs. These are micro-level assumptions related to Node B. Furthermore, the game tree includes assumptions about macro-micro transitions that are summarized by the vertical Arrow 1 in the diagram. The tree shows how an actor's payoff depends on the behavior of the other actor and vice versa, i.e., how the actors are interdependent. Next, rationality assumptions like the assumption of equilibrium behavior are represented by Arrow 2. Game-theoretic analysis yields propositions on equilibria of the Trust Game as well as the repeated version of the game and on properties of these equilibria. One such proposition is that the Trust Game has a unique subgame-perfect equilibrium in which trust is not placed and placed trust would be abused. The core proposition for the repeated game is that it has cooperative equilibria if and only if condition (1) is fulfilled. This permits us to derive implications concerning the behavior of rational actors, or in our case, implications concerning trustfulness and trustworthiness. These implications are represented by Node C. Finally, one can derive propositions on macro-level effects, for example that the outcome of the game with no trust placed leaves both actors worse off than another outcome that they could have attained, namely, the Pareto-optimal outcome of placed and honored trust.

Testable implications on embeddedness effects in the sense of qualitative predictions can then be derived by employing comparative statics with respect to condition (1). Rather than claiming that actors indeed follow the very strict behavioral rules of conditional behavior described above, we can use condition (1) to derive more qualitative predictions about behavior in the indefinitely repeated Trust Game. One proceeds from the observation that condition (1) is a necessary and sufficient condition for equilibria in the indefinitely repeated Trust Game such that trust is placed and honored throughout the game. One then assumes that placing and honoring trust as outcomes on the micro-level and, therefore, Pareto-optimality of the macro-outcome of the Trust Game, become more likely when condition (1) becomes less restrictive. This leads directly to testable predictions 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 with the temptation $(T_2 - R_2)/(T_2 - P_2)$ for the trustee. For our co-authors, the implication would be that Adam is more likely to accept Eve's first-authorship, *ceteris paribus*, when he expects to interact with Eve more frequently and for a longer time in the future, perhaps because Eve likewise has a tenured position. We would then expect that this increases the likelihood both of Eve investing time and effort and of a high-quality paper being produced.

The same approach to generating testable qualitative predictions from game theoretic models can be used to make further predictions on effects of macro-conditions (see Raub and Voss 1986: 316–321). One can specify how macro-conditions affect various features of the indefinitely repeated Trust Game, and how such conditions affect the likelihood of a cooperative equilibrium as a solution of the game. For example, placing and honoring trust by rational actors in the indefinitely repeated Trust Game is a result of implementing conditional strategies. Implementing such strategies presupposes observability of the partner's behavior. Thus, macro-conditions fostering observability will positively affect the likelihood of placing and honoring trust. Also, the folk theorem implies that placing and honoring trust in the indefinitely repeated Trust Game presupposes that rational actors need to somehow coordinate on a cooperative equilibrium (see also Voss 2015). Macro-conditions facilitating coordination include communication opportunities for the actors. They also include the availability of unambiguous rules on implementing a cooperative equilibrium and thus on placing and honoring trust conditionally, given that the conditions for cooperative equilibria are fulfilled. An important additional macro-condition facilitating trust are leaders consistently communicating such rules. For the co-authors, we would expect, for example, that successful collaboration is more likely if they are members of the same Department, with offices located close to each other and Department routines such as working in one's office rather than at home, as well as an overall cooperative "culture" supported by the Head of the Department that facilitates solving the coordination problem associated with establishing and maintaining conditional collaboration.

A remark on group size effects

A macro-condition for cooperation in social dilemmas other than the Trust Game that has received much attention in theoretical as well as empirical work is group size. Olson's (1965) analysis of group size effects on collective good production sparked much of this work. Game-theoretic analyses of group size effects can focus on non-repeated as well as repeated social dilemmas. For example, Diekmann (1985) derived predictions of group size effects in the Volunteer's Dilemma. He analyzed how increasing group size affects micro- as well as macro-outcomes. The micro-outcome is each actor's individual probability of providing the collective good in the symmetric equilibrium in mixed strategies. The macro-outcome is the associated probability that the collective good will be provided because at least one actor is willing to bear the costs. Diekmann shows that the game-theoretic model implies that increasing group size affects both probabilities negatively, while experimental evidence suggests empirical support only for the prediction with

respect to the micro-outcome.[28] For repeated social dilemmas such as an *n*-actor Prisoner's Dilemma or a Public Goods Game, one can analyze how increasing group size affects the condition for cooperative equilibria, namely, that the shadow of the future is large enough relative to the short-term incentives of the actors (see Raub 1988). Another prediction of group size effects in repeated *n*-actor social dilemmas follows when one takes into account that cooperative equilibria require each actor to receive reliable information on each other actor's behavior in each round of the game. This assumption will often be less likely to hold in games with many actors (see, e.g., Bendor and Mookherjee 1987).

Reciprocity: network control effects

Models of repeated Trust Games can be extended to account for control effects due to network embeddedness in addition to dyadic embeddedness. One then considers "games on networks" (Goyal 2007: Chapter 3; Jackson 2008: Chapter 9). In these extended models, the trustee interacts with a set of trustors, while the trustors are connected through a network that allows for communication about the behavior of the trustee. The focal Trust Game is now embedded in a more complex game that comprises Trust Games between the trustee and different trustors. The important feature is that the trustor in a focal Trust Game can transmit information on the trustee's behavior in that game to other trustors. In addition to the direct reciprocity exercised by the trustor who interacts with the trustee in the focal Trust Game, network embeddedness allows for indirect reciprocity exercised by other trustors. A trustee contemplating whether to honor or abuse trust in a focal Trust Game now must consider future sanctions by the trustor with whom he interacts in the focal Trust Game, as well as sanctions that can be applied by other future trustors who receive information on the trustee's behavior in the focal Trust Game and who may condition their future behavior on that information. In terms of our co-author example, Adam must take into account the consequences of how he collaborates with Eve for the future behavior of other colleagues with whom he collaborates in other joint projects or in teaching.

First, such network embeddedness can be a substitute for dyadic embeddedness (see Kreps 1990a: 106–108). Assume that the trustee interacts with a different trustor in each round of the indefinitely repeated Trust Game. Thus, each trustor plays the Trust Game only once with the trustee. Dyadic embeddedness is then removed

[28] The relation between the micro- and the macro-level probability is not trivial. Since it is enough if a single actor bears the cost of collective good production in the Volunteer's Dilemma, the negative group size effect on the individual probability could be "compensated" by an increase in the number of actors who may decide to produce the collective good.

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, the trustors as a "team" can condition their behavior in the same way as a trustor who plays each round. Again cooperative equilibria exist if and only if condition (1) is fulfilled. Hence, we see that network embeddedness can induce trust among rational and self-regarding actors.

Dyadic as well as network embeddedness is included in more complex models. Raub and Weesie (1990) analyze what is likely to be the first game-theoretic model of network effects for a social dilemma, namely the Prisoner's Dilemma (see Rapoport, Diekmann, and Franzen 1995 and Corten et al. 2016 for experimental evidence). Buskens (2002: Chapter 3) provides a related model of network effects for repeated Trust Games. In Buskens' model, a trustee interacts with a trustor in an indefinitely repeated Trust Game. After the interaction with a given trustor ends, the trustee goes on playing an indefinitely repeated Trust Game with another trustor, while information on the trustee's behavior in the Trust 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. The model is relatively general and allows for condiderable heterogeneity with respect to various features: the incentive T_2 for abusing trust varies between games, the probability of interactions starting with the trustee as well as the continuation probability for these interactions varies between trustors, and the probability of information transmission vary between pairs of trustors. One can then study subgame-perfect 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. A nice feature of these models is that they account for the intuition that trust will not always be placed. In addition to predictions concerning the likelihood of trust being affected by the shadow of the future and the trustee's short-term incentives, such models also allow us to predict the effects of network characteristics. Specifically, the likelihood of placing and honoring trust in a focal Trust Game increases with the density of the network of trustors as well as with the trustor's outdegree, that is, the probability of her transmitting information to the next trustor who interacts with the trustee. This is intuitively plausible since network density as well as outdegree increase the trustor's sanction possibilities. Hence, if the trustee considers the long-term consequences of his behavior, higher network density and outdegree foster the placing and honoring of trust even if the trustee's short-term incentive to abuse trust is fairly large.

In our co-author example, assuming that Adam and Eve are members of the same Department, we would expect that they will be *ceteris paribus* more likely to cooperate productively and to come up with a high-quality paper if two kinds of conditions are fulfilled simultaneously. The first is that Adam is likely to collaborate

in research as well as teaching with several other members of the Department in the foreseeable future. The second is that Eve is in contact with those other members too, through her own collaborations with colleagues or through other informal interactions.

One problem of these models of network effects is that they assume that information is reliable and that incentive problems associated with the supply of information are neglected (see, e.g., Lorenz 1988; Raub and Weesie 1990: 648; Williamson 1996: 153–155; Blumberg 1997: 208–210; Buskens 2002: 18–20). However, supplying information on the trustee's behavior is a contribution to a public good, namely, enforcing trustworthy behavior on the part of the trustee. Such contributions are problematic: after all, public good production is itself a social dilemma when contributions are costly (this feature is often discussed as a major problem of institutions such as eBay's feedback forum; see, e.g., Bolton and Ockenfels 2009). Moreover, information from third parties can be inconsistent with one's own experiences. Information can also be problematic owing to misunderstandings or strategic misrepresentations: imagine that the trustors are competitors who purchase the same goods from the same seller. In a nutshell, one would expect that the effects of network embeddedness are attenuated when such problems become more serious.

Models such as Buskens' address network control in the sense that other trustors can sanction the trustee in future interactions. This is control through "voice" (Hirschman 1970). A different case of network control is that a trustor has access to alternative trustees and can exercise control through "exit": whether or not the trustor interacts again with the trustee in the future depends on the trustee's behavior in the focal Trust Game. Modeling network control through exit opportunities for the trustor is not trivial (see Hirshleifer and Rasmusen 1989; Schüssler 1989; Vanberg and Congleton 1992 for related models) but one would generally expect that the likelihood of placing and honoring trust increases with the trustor's exit opportunities.

It should be noted that, rather than fostering trust and cooperation, network embeddedness can also have adverse effects for the actors who are involved in social dilemmas. Focusing on learning and information diffusion rather than game-theoretic rationality as a driving force of behavior, Burt and Knez (1995) have shown that dense networks can amplify trust as well as distrust. The core argument is that due to the homogeneity of opinions in a dense network, actors become convinced of some information because they receive the information disproportionately often. Co-ethnics may be able to solve trust problems in economic exchange by transacting with each other, but this may lead to entrapment and to their missing opportunities from outside networks (e.g., Portes 1998). Flache (2002) offers a game-theoretic model of how informal social ties between the members of a team can undermine cooperation between the members of the team because they have

to trade off the benefits of sanctioning team members who do not cooperate against the costs of deteriorating informal social ties due to negative sanctions. While there is considerable empirical research on adverse effects of embeddedness, systematic theoretical modeling of such effects is scarce.

Learning effects through embeddedness

The game-theoretic models of control effects on trust through embeddedness employ (repeated) games with complete information: roughly speaking, each actor is informed about the behavioral alternatives and incentives of all actors. Specifically, trustors have complete information about the trustee's behavioral alternatives and incentives. Hence, there is no need – and no opportunity – for trustors to learn about the trustee's unobservable characteristics during the game. This means that these models do not yield predictions concerning the learning effects of embeddedness.[29]

Predictions concerning control and learning effects can be derived from models of games with incomplete information. Typically, these are models of finitely repeated games. To get a flavor of these games, consider first of all a finitely repeated game with complete information. Assume that trustor and trustee play the Trust Game in Figure 2 N times. Equilibrium behavior clearly requires trust to be abused and no trust to be placed in the final round. Therefore, behavior in the penultimate round cannot affect behavior in the final round. Hence, no trust will be placed in the penultimate round and so forth, back to the first round. This is the famous backward induction argument showing that placing and honoring trust cannot be a result of rational and selfish behavior in a finitely repeated Trust Game with complete information.

Things change radically by introducing *incomplete information in the finitely repeated Trust Game* (see Camerer and Weigelt 1988; Dasgupta 1988; Buskens 2003). Assume once again that there is a positive ex ante probability π that the trustee actually has no incentive to abuse trust, i.e., his payoff from abusing trust is $T_2^* < R_2$ (again, an alternative assumption leading to essentially the same results would be that the trustee, with probability π, has no opportunity to abuse trust). The trustor knows the probability π but cannot observe directly whether the trustee's payoff from abusing trust is T_2^* or T_2. Now, if the trustor places trust in some round of the repeated game that is not the final round, trust may be honored for one of two very different reasons. First, the trustee's payoff could be $T_2^* < R_2$ so that there is no incentive at all for the trustee to abuse trust. Second, the trustee's payoff could be $T_2 > R_2$ but

[29] One might argue that learning is still possible in these models, since there are many equilibria and it is not clear why actors should choose the same equilibrium to start with. We disregard this issue, assuming that actors coordinate instantly on the same equilibrium (see, e.g., Fudenberg and Levine 1998: 20).

the trustee has an incentive for reputation building. The trustee knows that if he abuses trust, the trustor can infer for sure that the trustee's payoff from abusing trust is $T_2 > R_2$ and she will thus never place trust again in future rounds. On the other hand, if the trustee honors trust, the trustor remains uncertain about the trustee's incentives and she may place trust again in the future. Conversely, the trustor can anticipate such behavior on the part of the trustee and may therefore be inclined to indeed place trust. In this game, the trustor can control the trustee in the sense that placing trust in future rounds will depend on honoring trust in the current round and the trustor can learn about the trustee's incentives from his behavior in previous rounds. The result is a subtle interplay between a trustor who tries to learn about and 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 the game has a sequential equilibrium (a further refinement of the subgame-perfect equilibrium concept that can be applied to games with incomplete information, Kreps and Wilson 1982) that does involve placing and honoring trust in some rounds of the repeated game. More precisely, in that equilibrium, the game starts with trust being placed and honored in a number of rounds. Afterwards, 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 equilibrium behavior can induce considerable honored trust even if the probability π that the trustee has no incentive to abuse trust is small. In the equilibrium, learning occurs – in the sense that the trustor updates her belief about the probability that she is playing with a trustee without an incentive to abuse trust – if trust is abused and in the second phase as long as trust is honored. Learning is rational in the sense of Bayesian updating. The first phase of the game where trust is 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 ex ante probability π. While the trustor's risk is a driving force of the model, the trustee's temptation $(T_2 - R_2)/(T_2 - P_2)$ only affects behavior in the second randomization phase of the repeated game.

Game-theoretic models with incomplete information like the finitely repeated Trust Game are complex and not easily analyzed. They become even more complex by including learning due to network embeddedness. A shortcut linking learning effects of network embeddedness to such models would be to assume that the trustor's ex ante probability π of interacting with a trustee who would never abuse

trust depends on information the trustor receives from third parties, such as other trustors who played Trust Games previously with the trustee. Specifically, based on information diffusion models in networks of trustors (e.g., Buskens 2002: Chapter 4) and assuming that the information about the trustee is positive (it shows that the trustee has honored rather than abused trust), one would expect the ex ante probability π to increase with the density of the network of trustors as well as with the extent to which the trustor in the focal Trust Game receives information about the trustee from other trustors, i.e., increases in the trustor's indegree.

A more explicit game-theoretic model of network effects in games with incomplete information has been provided by Buskens (2003). In that model, the trustee plays Trust Games with two different trustors *A* and *B*. With some probability, each trustor can inform the other trustor about the trustee's previous behavior. We can conceive of the probability that trustor *A* transmits information to trustor *B* as *A*'s outdegree and *B*'s indegree (and vice versa). Thus, trustor *A* controls the trustee through her outdegree and learns from *B* about the trustee through her indegree. 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 repeated game in which trust is placed and honored becomes longer and in this sense network embeddedness increases trust.

Summarizing and interpreting the results of the game-theoretic models for learning effects through dyadic and network embeddedness yields the predictions that the likelihood of placing and honoring trust decreases with the trustor's risk $(P_1 - S_1)/(R_1 - S_1)$ and increases if the trustor's previous experiences with the trustee are positive (the trustee honored trust) rather than negative (the trustee abused trust). Furthermore, assuming that the trustor receives positive information about the trustee from other trustors, the likelihood of placing and honoring trust increases with the density of the network of trustors, and with the trustor's indegree.

Models for control and learning effects of embeddedness in games with incomplete information are not only problematic in that they use very strong assumptions about the actors' rationality (in the sense of sequential equilibrium), including rational (Bayesian) updating of beliefs. These models are also problematic in that they neglect learning regarding features other than the unobservable characteristics of the trustee. For example, a trustor could try to use information she receives from other trustors to infer how to reasonably cope with trust problems. Also, past interactions may give rise to effects other than exclusively learning. For example, actors may have pledged investments in their relationship through past interactions and these investments affect the incentives in the focal Trust Game (Williamson 1996; Batenburg, Raub, and Snijders 2003).

The most attractive feature of game-theoretic models involving incomplete information is that control and learning can be analyzed simultaneously. The price tag attached to these models is a set of rather strong assumptions about the actors' rationality. Alternatives are "pure" learning models in which actors adapt their behavior based on past experiences. Actors try to optimize short-term outcomes while not (or "hardly") looking ahead. This implies, too, that actors do not take other actors' incentives into account (see, e.g., Camerer 2003: Chapter 6 for a useful overview of learning models; Macy and Flache 1995, 2002, and Flache and Macy 2002 provide applications to social dilemmas; Buskens 2002: Chapter 4 is an example of a model of learning in networks). Hence, these models neglect control effects. Typically (see Buskens and Raub 2002: 173–176), learning models yield predictions stating that the likelihood of placing trust decreases with the trustor's risk $(P_1 - S_1)/(R_1 - S_1)$. Also, the trustor's estimation of the probability π that trust will be honored will typically increase with positive information about the trustee's behavior in previous interactions, be it information concerning the trustor's own previous interactions with the trustee or information from third parties. Therefore, one would again predict that more positive information increases the likelihood of trust being placed.

Rational models for running an academic institution

Social dilemma-like situations and cooperation problems are pervasive in academic institutions. Collaboration between academics in joint research projects, as co-authors, and in co-teaching often comes with incentive problems that are typical of social dilemmas. Departments, research groups, and teams jointly running a teaching program benefit from an intellectual "climate" that includes mutual support as well as constructive criticism of each other's work and the willingness to learn from such criticism. The "climate" includes collaboration between academic and support staff (in Dutch: *OBP*), with support staff supporting teaching as well as research and academic staff respecting these contributions and knowing how to make the best use of the expertise that support staff can provide. Creating and maintaining an intellectual climate of this kind is similar to producing a public good because such a climate depends on the individual contributions of the academics involved and because contributions come with costs in terms of time and effort and thus incentives for free riding.

Other cooperation problems arise when leaders of academic groups have to agree on how to spend common resources or on common policies. For example, the chair holders within a Department have a collective interest in agreeing on how to run

the Department and on how to spend resources, if only because this strengthens their bargaining position vis-à-vis the Dean. However, chair holders likewise have an individual interest in maximizing resources for their own chair. Heads of Departments sometimes face similar problems when interacting with each other, as do Deans interacting with each other and with the university's Executive Board.

Another example is what could be labeled the "quality-quantity dilemma" which was discussed a while ago within our university (e.g., Raub 2013). This dilemma is related to a reward system in science that focuses on quantitative indicators such as numbers of publication, specifically in journals with a high impact factor, the H-index, and so forth (roughly: substituting counting for reading publications). This reward system induces incentives for strategic behavior such as "salami slicing" in publication behavior: maximizing the number of publications by publishing papers with the minimum amount of information and possibly many co-authors. Such publication behavior does not necessarily serve the growth of knowledge.

Finally, limited resources – a core feature of academic life – can easily induce policies that create strategic interdependencies between academics and, as a possibly anticipated but unintended by-effect, can also induce cooperation problems. For example, limited resources imply that it is often not feasible to hire individual "star researchers" and that growth of knowledge is more dependent on high-quality teamwork, while teamwork comes with cooperation problems.

These examples raise the question of what can be learned from models of trust and cooperation in social dilemmas when it comes to running an academic institution.[30] Robust answers that do not depend on very specific model assumptions can be derived by considering that, first, embeddedness is often an important feature of interactions in academia. Second, those running academic institutions often have opportunities to actively shape embeddedness.

Interactions in academic institutions are typically embedded in long-term relationships and in networks so that cooperation can, in principle, result from equilibrium behavior and a self-enforcing norm of conditional cooperation. At the same time, due to the folk theorem and the equilibrium selection problem, academics

[30] I consider policy measures that facilitate cooperation between purely self-regarding actors. Note, first, that value judgments should be applied with care in our context. Consider a group of scholars who fail to establish and maintain cooperation for a joint paper or in co-teaching. The point is *not*, at least not necessarily, that they are idling. Rather, they may be working hard on other research projects or on other courses. Second, we need *not* assume that scholars, in general, are purely self-regarding (casual observation suggests that some of them are quite consistently not self-regarding, while many are at least sometimes not self-regarding). Rather, the policy measures discussed can be useful to facilitate cooperation *even if* everybody *were* purely self-regarding (see Voss 2016). What I do assume, though, is that scholars, like other human beings, are "purposeful" and do react systematically to incentives.

face the problem of coordinating on a cooperative equilibrium. Hence, when the conditions for conditional cooperation and reciprocity are favorable, those responsible for running an academic institution should clearly and unambiguously communicate that cooperation is feasible in the sense of being equilibrium behavior. They should also focus on communicating and supporting rules that make cooperative equilibria, if they exist, the anticipated solution to cooperation problems. Thus, consistency and predictability with respect to policy become important for academic leaders (see Baron and Kreps 1999). For example, stressing both the importance of a long-term perspective and clear general rules of behavior, such as "cooperate conditionally whenever this is in everybody's enlightened self-interest," and the disadvantages of opportunistic short-term policies should go hand in hand with avoiding tête-à-tête management (in Dutch: *bilateraaltjes*) and management by "divide and rule," precisely because both types of management focus on short-term results.

Those responsible for running academic institutions are often able to influence the incentives for cooperation. Technically speaking, it is often possible to affect the payoffs associated with cooperation and defection in social dilemmas, at least to some degree. For example, consider reward systems that focus not only on individual achievements but also on team performance. Through a reward system that accounts explicitly or at least *de facto* for team performance, say, the results of external evaluations of a Department's research and teaching programs, one can increase individual rewards R_i for mutual cooperation as well as decrease individual rewards P_i when cooperation fails. This affects individual risks of cooperation and the individual temptation to defect so that the conditions for a cooperative equilibrium become less restrictive and, hence, cooperation becomes more likely. Note that reward systems with a focus on team performance may even contribute to mitigating the "quality-quantity dilemma" since groups are in an at least somewhat better position than individuals to cope with this dilemma (see Raub 2013). Similarly, it might be useful to have a long-term policy that invests in good support for researchers preparing grant proposals from a Research Support Office and by having peers provide systematic feedback on draft proposals, rather than investing in individual "star researchers."

We have seen that network embeddedness can affect cooperation through learning. What happens in interactions between two actors can function as a signal for third parties and affect their future behavior. This can be taken into account when running an academic institution. For example, those running such an institution have an interest in researchers generating new funding for a research project. Such funding is not only beneficial for the applicant and principal investigator (PI), but it also contributes to the overall quality of the PI's Department, is valuable for the Faculty in several respects, and can be seen as a contribution to collective goods. On the other hand,

carefully preparing applications requires creativity as well as time and effort, with a low probability of success, certainly in the case of prestigious and very competitive funding schemes. Under these circumstances, rewarding success, for example through promotion or tenure, not only affects the individual PI. Through network embeddedness, other researchers and potential applicants will typically come to know that success tends to be rewarded. Such rewards therefore function as credible signals for other researchers that own success will also be rewarded, thereby stimulating them in their efforts.

The measures discussed so far take embeddedness as given. However, the embeddedness of interactions in academic institutions need not always be taken as given. Those responsible for running an academic institution can sometimes modify embeddedness characteristics so that cooperation becomes more likely. For example, consider the time horizon of postdocs and faculty. We have seen that a longer time horizon (technically speaking: increasing the continuation probability w) increases the likelihood of cooperation of even purely self-regarding if actors. Of course, it is not feasible and for various reasons also not sensible, even if it were feasible, to offer everybody tenure. However, given the choice between, say, hiring four postdocs for one year each and one postdoc for four years, it follows from rational models of trust and cooperation in social dilemmas that it makes sense, *ceteris paribus*, to hire one postdoc for four years, thus inducing more incentives to contribute to collective good production rather than incentives to exclusively focus on short-term individual gains. Also, and in the same spirit, when it comes to collaborative projects or co-teaching, cooperation will be facilitated by seeing to it that the same colleagues are involved repeatedly in collaborative projects or in co-teaching, and possibly also in collaborative research projects *as well* as in co-teaching, since this fosters opportunities and incentives for conditional cooperation and reciprocity.

Likewise, our analysis revealed that and why it is important for actors to observe and monitor each other's behavior. It immediately follows that it makes sense to provide a shared location for the members of a team such as a Department, including a shared location that is attractive as a workplace and thus mitigates incentives to work individually elsewhere. Given that many colleagues prefer a small office, to be shared with at most a very limited number of colleagues, to an open-plan office (in Dutch: *kantoortuin*) and given that working at home will not facilitate contact with colleagues per se, be aware of unintended side-effects of what is known as "flexible, mobile, and remote working" (in Dutch: *het nieuwe werken*) at least in an academic context.

Rational models of cooperation in social dilemmas shed light on size effects. Very often, increasing group size has a negative effect on conditional cooperation and reciprocity. This has policy implications, too, for academic institutions. It provides a rationale, for example, for organizing academic institutions in a way that curtails the size of groups in which members are confronted with cooperation problems

similar to social dilemmas. *Ceteris paribus*, two colleagues co-teaching a course can cooperate more easily than a much more sizeable group of teachers. Hence, be aware of the risks of what is known in Dutch as *docentenparade* (see Raub 2014 for a game-theoretic model of co-teaching). See to it that, say, a Faculty is organized into subgroups that are not too large so that they offer good conditions for conditional cooperation between the group members. See to it that such subgroups belong to larger units, with the subgroups as actors in the larger units, and that these larger units are small enough to offer good conditions for conditional cooperation between the subgroups, and so forth (see, for example, Olson 1965: 62 on "a federal group that is divided into a number of small groups"). Of course, following this policy once again requires those responsible for the subgroups ("academic leaders") to understand and implement the logic of conditional cooperation.

Group size effects are important, too, because these effects highlight that fostering conditions for cooperation must be balanced with other aims of running academic institutions. For example, while stable small groups are good contexts for self-enforcing conditional cooperation, they also come with risks, such as loss of privacy for individual group members and strong within-group solidarity that inhibits between-group cooperation. Perhaps most important in an academic context, stable small groups may inhibit innovativeness (see, e.g., Coleman 1990: Chapter 5). Therefore, those responsible for running academic institutions have reasons to focus on groups of intermediate size, for example, a Department (in Dutch: an *afdeling* or a *departement*) as the core groups in which cooperation should be fostered, rather than on very large groups such as a Faculty as a whole or very small groups such as individual chairs and its members. Subsequently, furnish opportunities and incentives for cooperation in groups of intermediate size, including sufficient autonomy for such groups. This includes degrees of freedom with respect to spending a lump-sum budget. It also includes increasing that lump-sum rather than retaining large portions of the budget for separate projects on the Faculty level. *Facultaire potjes*, as these are known in Dutch, should preferably be reduced when aiming to foster cooperation within groups that make up the Faculty rather than run the Faculty like a planned economy. Likewise, see to it that groups of intermediate size create good contexts for innovation and fresh ideas by fostering diversity across various dimensions, including open and external selection procedures for filling vacancies.

Note that the policy measures sketched here also illustrate that those running an academic institution can target two kinds of effects on the quality of research and teaching and, therefore, on the growth of knowledge. First, they can target direct effects. For example, assembling a research team or a small group of teachers for a course so that they have complementary expertise will likely have a direct positive effect on the likelihood of a new research finding or a well-taught course. Second, they can target

an indirect effect. Complementary expertise between the members of a team might likewise have an incentive effect on cooperation, because cooperation yields more attractive results, given complementary expertise. The indirect effects of policy measures via incentives for cooperation should not be overlooked in academic institutions.

2.2 Empirical research on embeddedness effects[31]

In the following, I provide a brief overview of evidence of embeddedness effects from a number of empirical studies on trust problems. These are studies within and resulting from the NWO PIONIER-program *The Management of Matches*. I was fortunate to have had the opportunity to collaborate on these studies not only with Jeroen Weesie and Vincent Buskens but also with colleagues such as Frits Tazelaar, Chris Snijders, and Ronald Batenburg.

A sizable literature covering empirical research on trust problems and, more specifically, on embeddedness effects on trust is now available. This includes experimental literature and field research (see Buskens and Raub 2013 for an overview). Much of this literature offers evidence for effects of embeddedness, although we can seldom determine whether the effects are due to learning, control, or a combination of the two mechanisms. One aim of our own studies has been to disentangle the different mechanisms through which embeddedness affects trust. The other aim has been to use alternative and complementary research designs – survey research, vignette studies, and lab experiments – to repeatedly test basically the same predictions of embeddedness effects. Table 2 summarizes the predictions of embeddedness effects, while Table 3 provides an overview of key features of the four studies.

Frits Tazelaar

Chris Snijders

Ronald Batenburg

[31] The sketch of empirical work uses materials from Raub and Buskens (2008). See Raub and Buskens (2008) as well as the representative publications mentioned for details and complications that I have chosen to omit in this overview.

Table 2: Predictions of embeddedness effects.

Two mechanisms	Two types of embeddedness	
	Dyad	Network
Control	Trustfulness and trustworthiness decrease with the trustee's temptation and increase with the shadow of the future.	Trustfulness and trustworthiness increase with the density of the trustor's network and her outdegree.
Learning	Trustfulness and trustworthiness decrease with the trustor's risk and increase (decrease) with positive (negative) experiences with a trustee.	Trustfulness and trustworthiness increase with the density of the trustor's network and her indegree (given that information about the trustee is predominantly positive).

A survey study on buyer-supplier relations

Trust in economic exchange is the topic of a survey study on the purchase of information technology (IT) products – hardware and software – by Dutch small- and medium-sized enterprises (SMEs; 5–200 employees; see Batenburg, Raub, and Snijders 2003; Buskens, Raub, and Weesie 2000; Rooks, Raub, and Tazelaar 2006 for representative publications).[32] In this study, key informants of the buyer firms, typically the IT managers responsible for the purchase, provided information on the purchase of IT products, using a structured questionnaire. The study comprises data from almost 800 buyer firms on somewhat more than 1200 transactions. We conceptualize an IT transaction as a variant of a Trust Game, with the buyer in the role of the trustor and the supplier in the role of the trustee.

The survey allows for measuring the trustor's (lack of) trustfulness through her investments in the ex ante management of the transaction. This refers to the buyer's investments in negotiating and contracting with the supplier such as the number of person-days of the buyer's employees that have been involved in negotiating and contracting, the number of the buyer's departments that have been involved, the use of external legal advisors, the use of a standard contract or a tailor-made contract, and, finally, the number of financial and legal clauses as well as technical specifications that have been addressed during the negotiations or that are included in the contract. Such ex ante management reduces the incentives of the supplier

[32] Blumberg (1997) is a related study on Research & Development alliances.

Table 3: Overview of the empirical studies.

	Survey	Vignette study 1	Vignette study 2	Lab experiment
Type of problematic social situation	Buyer-supplier transactions (TAs)	Buyer-supplier TAs	Buying a used car	Trust Games (TGs)
Respondents/subjects	IT-managers	Purchase managers	Students	Students
Number of firms/ respondents/subjects	788	40	125	72
Number of observations	1252 TAs	348 vignettes/TAs	1249 pairs of vignettes/TAs	2160 TGs
Dependent variables				
Trustor	Investments in ex ante management of TAs	Investments in ex ante management of TAs	Choice between two dealers	Placing trust
Trustee	Performance	---	---	Honoring trust
Independent variables				
Temptation, risk	Various TA characteristics	Various TA characteristics	Price of the car	Payoffs not varied
Dyadic embeddedness				
Learning	(Satisfaction with) own previous TAs with supplier	(Satisfaction with) own previous TAs with supplier	(Satisfaction with) own previous TAs with dealer	Trustee behavior in own previous TGs
Control	Expected future TAs with supplier	Expected future TAs with supplier	Whether or not buyer expects to move	Rounds left in the repeated game
Network embeddedness				
Learning	Degree, density, visibility of supplier	Degree	Third-party information, density	Trustee behavior in previous TGs with other trustor
Control	Degree, density, visibility of supplier, exit opportunities	Degree, exit opportunities	Density, degree	Information condition
Other variables	Size of buyer and supplier firm, in-house legal expertise of buyer, etc.	Respondent characteristics: experience etc.	Not applicable	Timing within the experiment
Representative publications	Batenburg, Raub, and Snijders 2003; Buskens, Raub, and Weesie 2000; Rooks, Raub, and Tazelaar 2006	Rooks, Raub, Selten, and Tazelaar 2000	Buskens and Weesie 2000	Buskens, Raub, and Van der Veer 2010

to engage in opportunistic behavior and likewise mitigates the damage for the buyer in the event of supplier opportunism. For example, ex ante management of the transaction yields contractual agreements on compensation for the buyer in case of delivery delays or quality problems. At the same time, ex ante management is associated with costs for the buyer in terms of time and money. Therefore, the assumption is that the less the buyer trusts the supplier, the more she will invest in ex ante management.

The measurement of (lack of) trustworthiness of the supplier uses ex post problems after the transaction. Examples of indicators in the survey for such problems are delivery delays, exceeding the agreed upon price or budget, various quality problems, and after sales problems such as insufficient support or service.

Independent variables related to the supplier's temptation to behave opportunistically and the buyer's risk associated with the transaction are transaction characteristics such as the volume of the transaction, the buyer's ex ante monitoring problems with respect to the quality of the product, the buyer's switching costs in case the product had to be replaced, and the importance of the product for the buyer, for example in terms of the buyer's profitability.

The survey includes variables representing the dyadic embeddedness of the transaction. With respect to previous experiences of the buyer with the supplier, respondents are asked whether they have transacted any previous business with the supplier and, if so, how satisfied they were with this previous business. Information on the shadow of the future is obtained by querying the buyer's expectations of the frequency and size of future business with the supplier at the time of the focal transaction.

Finally, the survey makes it possible to measure the network embeddedness of the transaction. First, information is available on the buyer's network with other clients of the supplier, i.e., the buyer's voice network. More precisely, the survey includes an indicator for the degree of the buyer in the sense of the number of other clients of the supplier whom the buyer knows. Furthermore, an indicator is available for the density of contacts between firms in the buyer's sector of industry. Another indicator for the density of the network in which the transaction is embedded is the geographical distance between the municipalities in which buyer and supplier are located, assuming that the buyer will typically have more contacts with other buyers of the supplier and that these other buyers will have more contacts among themselves when buyer and supplier are located closer together. The opportunities for voice of buyers are indicated by a survey question on the visibility of the supplier in the market. Each of these network variables is related to the buyer's opportunities for learning and control through network embeddedness: they do not allow network learning and network control effects to be disentangled. In addition to indicators for

the buyer's voice network through contacts with other buyers, there are indicators available for the buyer's exit network. These include the number of potential suppliers of the product and the number of alternative products. These indicators are related to the buyer's opportunities to control the supplier through network embeddedness.

In addition, the survey provides information that makes it possible to control for such variables as the size of buyer and supplier firm and in-house legal expertise available in the buyer firm.

A vignette study on buyer-supplier relations

A second study (see Rooks, Raub, Selten, and Tazelaar 2000 for a representative publication) is a vignette study on how embeddedness affects trust (see Auspurg and Hinz 2015 on vignette designs). This study parallels the survey on IT-transactions in important respects. The focus is again on buyer-supplier relations and on the buyer's investments in the ex ante management of transactions. Various variables representing transaction characteristics and the embeddedness of the transaction closely resemble variables from the survey. Thus, while the survey and the vignette study employ very different designs, very similar predictions can be tested.

In this vignette study, the subjects are purchase managers at Dutch medium-sized and large companies. The managers are presented with hypothetical transactions and are asked to imagine that these *vignettes* are actual transactions from their daily business practice. The subjects then answer questions about their behavior related to these transactions. Once again, we conceptualize a transaction as a trust problem, with the buyer in the role of the trustor and the supplier in the role of the trustee. The study comprises data from 40 managers who provided judgments on almost 350 vignettes, 8–10 per respondent.

A vignette consists of a description of a transaction for which eight variables are varied. Three variables are related to transaction characteristics and can be used as indicators for the supplier's temptation to engage in opportunistic behavior and the buyer's risk associated with the transaction. The variables include the volume of the transaction, the extent to which the buyer can judge the quality of the product at the time of the transaction, and the extent to which the buyer needs to make specific investments to implement the transaction. These variables are closely related to the transaction characteristics measured in the survey, with both studies comprising a variable measuring the volume of the transaction, variables measuring the buyer's monitoring problems, and variables measuring specific investments, operationalized in the survey as switching costs.

Two variables represent dyadic embeddedness and describe the bilateral relationship between the buyer and the supplier. One of these variables indicates

whether buyer and supplier have transacted business with one another before and how satisfactory or problematic earlier transactions with the supplier have been. This variable represents learning through dyadic embeddedness. The other variable indicates whether buyer and supplier expect to do more business with each other in the future. This variable represents control opportunities through dyadic embeddedness. Network embeddedness is likewise represented by two variables. A degree variable indicates the extent to which the buyer knows other business partners of the supplier. This variable represents learning and control through network embeddedness. Another variable refers to the buyer's exit opportunities by describing the extent to which the buyer can access alternative suppliers to purchase a similar product. This variable represents control through network embeddedness. Note that the variables representing dyadic embeddedness and network embeddedness are close analogues of the respective variables in the survey on buyer-supplier relations. A final variable in the vignettes indicates the country in which the supplier is located. The underlying idea is that trust is more problematic when the supplier comes from a different institutional setting. The design of the study also makes it possible to control for respondent characteristics such as years of experience as a purchase manager and the degree to which a vignette resembles the everyday practice of the respondent.

The dependent variable in this vignette study are investments in the ex ante management of the transaction described on the vignette, closely paralleling the respective variable in the survey. Just like in the survey, (lack of) trust of the buyer is measured through such investments. More precisely, subjects were asked how much time they would invest in negotiating and contracting the transaction, and which departments of their own company would be involved in these negotiations. This vignette study does not include information on actual supplier behavior associated with executing the transactions described in the vignettes. Thus, the study does not allow for testing predictions of embeddedness effects on trustee behavior.

A vignette study on buying a used car

A second vignette study (see Buskens and Weesie 2000 for a representative publication) addresses a classic example of a trust problem in economic exchange: buying a used car (Akerlof 1970). In this study, students are asked to compare pairs of situations for buying a used car whereby the properties of the relationship between the buyer and the car dealer are varied. Thus, students are assumed to have the role of the trustor, while the dealer is assumed to be in the role of the trustee. There are various differences between the two vignette studies. The type of transaction is clearly different. In the first study, subjects rate vignettes by indicating the level of investments in negotiations and contracting, whereas in the second study subjects

compare different vignettes rather than provide ratings. In the first vignette study, subjects are purchase managers with considerable experience in the type of transactions discussed in the experiment, whereas in the second study the subjects are students. In the first vignette study, transaction characteristics are varied, whereas they are kept constant in the second study. In the first vignette study, learning and control through network embeddedness, in the sense of contacts with other buyers, are not explicitly distinguished, whereas the second study makes an attempt to do so.

In the second vignette study, students are asked to compare situations for buying a used car. The students have been presented with pairs of vignettes describing a transaction and subsequently stated their preference for one vignette within each pair. The preference for a vignette is assumed to indicate that the buyer trusts the respective dealer more. The study does not provide information on the actual behavior of the dealer associated with the transactions described in the vignettes. Thus, this study again does not allow for testing predictions of embeddedness effects on trustee behavior. The experiment was carried out in the U.S. (Chicago) and the Netherlands (Utrecht and Tilburg). The study comprises data from 125 subjects on almost 1250 comparisons of pairs of vignettes, i.e., ten comparisons per subject.

Six variables are varied in the vignettes. The first variable indicates the price of the car. The price of the car is held constant within each pair of alternative vignettes. Consequently, the price cannot have a direct effect on the choices made by the subjects, but it may be that some embeddedness variables are more important for cheap cars than for expensive ones. In other words, the volume variable is added to test interaction effects of the size of the trust problem and embeddedness variables.

Five other variables represent embeddedness characteristics. Whether the buyer has bought a car from the dealer before and was satisfied, or never bought a car from the dealer, represents dyadic learning. Control opportunities of the buyer at the dyadic level are operationalized as whether or not the buyer expects to move to the other end of the country soon. Control is more difficult for a buyer if she moves and the shadow of the future is thus smaller. Moreover, the probability of the buyer having future transactions with the dealer is smaller if the buyer moves.

Concerning network embeddedness, a density variable differentiates between a dealer whose garage is or is not well known in the buyer's neighborhood. If more potential customers in the neighborhood know the dealer, the buyer probably knows more of these customers and there are likely more ties among these other customers. Therefore, learning about and controlling a well-known garage through the network of customers can be more effective than learning about and controlling a garage that is not well known. Third-party information is operationalized as whether or not the buyer has information from friends about their transactions with the garage. This variable represents learning through network embeddedness. Finally, the design

of the study includes a degree variable, operationalized as whether or not both the buyer and the dealer are members of the same sports team. This is a measure of the buyer's degree in the sense that the number of acquaintances the buyer and dealer have in common is expected to be larger if the buyer and dealer are members of the same sports team. Common membership provides the buyer with opportunities to control the dealer through reputational sanctions both in his business and as a team member. These sanctions can include discouraging others from buying from the dealer, but also social sanctions while engaging in sports team activities. A rational dealer should be concerned about these sanction opportunities of the buyer.

An experiment on embeddedness effects in finitely repeated Trust Games

Our final study is an experiment employing a design that closely reproduces a finitely repeated Trust Game (Buskens, Raub, and Van der Veer 2010). In this experiment, Trust Games are played in triads comprising two trustors and a trustee. The outcomes of the Trust Games are points that subjects earn. Subjects are paid for points they earn at the end of the experiment. Subjects play the Trust Game in repeated games of 15 rounds. Subjects are matched in triads, namely, one trustee and two trustors. Clearly, a triad represents a small network between the subjects. In each of the 15 rounds, the trustee plays one Trust Game with each of the two trustors. First, one of the trustors, say, trustor 1, plays a Trust Game with the trustee. After this Trust Game is finished, trustor 2, plays a Trust Game with the same trustee. This pair of games is played 15 times. In every round, trustors play one Trust Game and the trustee plays two Trust Games, adding to 30 Trust Games played per repeated game by one trustee.

All subjects, trustors and trustee, have complete information about the whole structure of the game, such as the number of rounds to be played, the payoff function of trustors and trustee, etc. The experiment employs two information conditions. In both conditions, the trustee is immediately informed about the trustor's move in the current Trust Game. Between conditions, the amount of information shared among the two trustors playing with the same trustee is varied. In the "no information exchange between trustors" condition, trustors do not share any information: each trustor only knows what happens in her own Trust Games with the trustee but is not informed about what happens in the games that the other trustor is playing with the same trustee. In the "full information exchange between trustors" condition, trustors playing with the same trustee share all information about each other's games. In this condition, as soon as a game has been played, whether that is the first or the second game in the round, both trustors receive information on the choices made in this game. Information is provided automatically via networked computers and is always truthful.

During the experiment, every subject played a repeated game three times, once as a trustee, once as trustor 1, and once as trustor 2. Each subject played all three repeated games in the same information condition. In between the three repeated games, the subjects were rematched to other subjects. Subjects were never rematched to their partners in a previous repeated game. This was likewise made common knowledge for all subjects.

In total, 72 subjects participated in the experiment, mostly undergraduate students from different fields. Four sessions were scheduled and 18 subjects participated in each session. Two sessions were played in the condition with no information exchange between trustors and two sessions in the condition with full information exchange. With four sessions, six triads per session, three repeated games of 15 rounds per subject, and each round comprising two Trust Games, $4 \times 6 \times 3 \times 15 \times 2 = 2160$ Trust Games were played in total. Trustee behavior was observed only in those games in which the trustor was trustful. There were almost 500 games in which there was no trust, leaving more than 1600 games in which the trustee's behavior was observed.

This experiment can be used to study how the trustor's trustfulness and the trustee's trustworthiness depend on the trustor's control opportunities due to embeddedness. Dyadic control depends on the number of rounds left in the repeated game (shadow of the future). Network control depends on the information condition. The experiment can likewise be used to study how the trustor's trustfulness depends on her learning about the trustee through embeddedness. Dyadic learning of the trustor results from the trustee's behavior in previous rounds of the game with that trustor. Network learning results from information – if any – that the trustor receives from the other trustor on the trustee's behavior in previous rounds vis-à-vis the other trustor. Note that payoffs do not vary between the Trust Games so that the experiment does not study the effects of the trustee's temptation and the trustor's risk on trustfulness and trustworthiness.[33]

Summary of findings on embeddedness effects

The empirical evidence from the four studies is summarized in Table 4. Findings are based on a variety of statistical models. Typically, alternative statistical models have been estimated for each study to ensure the robustness of findings.

[33] The *Management of Matches*-program led to a considerable amount of experimental work. Vincent Buskens has been a driving force for much of this work. Through his own Utrecht University High Potential Program *Dynamics of Cooperation, Networks, and Institutions*, a collaboration with Stephanie Rosenkranz, he managed to establish his *Experimental Laboratory for Sociology and Economics* in 2005, at that time possibly the very first state-of-the-art experimental lab for sociology in Europe.

Table 4: Overview of evidence for predictions related to embeddedness effects from the empirical studies.

	Survey	Vignette study 1	Vignette study 2	Lab experiment
Effects of temptation and risk on trustor behavior	Support	Support	No test	No test
Effects of temptation and risk on trustee behavior	Support	No test	No test	No test
Dyadic embeddedness				
Learning effects on trustor behavior	Support	Support	Support	Support
Control effects on trustor behavior	Support	Support	Support	Support
Control effects on trustee behavior	No support	No test	No test	Support
Network embeddedness				
Learning effects on trustor behavior	Hardly/ no support	Support	Support	Support
Control effects on trustor behavior	Hardly/ no support	Some support	Some support	No support
Control effects on trustee behavior	Support	No test	No test	Support

We find consistent evidence for the predictions of the effects of temptation and risk on trustor and trustee behavior, for the predictions concerning dyadic learning and control effects on trustor behavior, and for the predictions of network control effects on trustee behavior. Quite some, but not all, evidence supports the predictions of network learning effects on trustor behavior and the prediction concerning dyadic control effects on trustee behavior. However, while we do find much evidence for network control effects on trustee behavior, there is little evidence for network control effects on trustor behavior.

The latter pattern of findings – network control effects on trustee behavior but hardly any on trustor behavior – is challenging and arguably the biggest challenge that emerges from the results of the four studies. Buskens (2002: 152–161) provides various arguments and empirical evidence that the lack of effects of network control opportunities on buyer behavior in the data from the IT transactions survey is at least partly due to design, data, and measurement problems of the survey, including problems related to endogeneity of network embeddedness characteristics and sample selectivity. Clearly, these are no plausible arguments for the lack of network control

effects on trustor behavior in the lab experiment. Thus, one might wonder whether the findings concerning the effects of control opportunities through network embeddedness indicate limits of strategic rationality. First, consider the situation of the trustee (or, respectively, the supplier). He has a good reason to react to the trustor's dyadic control opportunities *as well as* her network control opportunities if he anticipates that his present trustworthiness might affect the future trustfulness of the same or other trustors. Similarly, the trustor has a good reason to react to her dyadic control opportunities if she anticipates that the trustee is considering how his present trustworthiness will affect her *own* future trustfulness. However, the trustor needs to reason "more steps ahead" before having a good reason to react to her network control opportunities. Namely, she must anticipate that the trustee is considering how his present trustworthiness will affect future trustfulness of *other* trustors *and* that *other* trustors will in fact condition their trustfulness on the trustee's present trustworthiness. It may be that actors are less likely to reason so many steps ahead, certainly in unfamiliar settings such as the lab experiment (see Binmore 1998: Chapter 0.4.2 for general arguments in this direction). For example, if the conjecture is correct, one would expect that effects of network control opportunities on trustor behavior will be easier to find when trustors play repeated Trust Games with information exchange between trustors many times and specifically when they also take on the role of the trustee in some of those repeated game. Note that the findings from the vignette study on buyer-supplier relations are in line with this reasoning. After all, it is the trustor behavior of experienced subjects, namely purchase managers, that is observed in that study and these subjects do indeed react to network control opportunities (see Buskens, Raub, and Van der Veer 2010: 311–312 and Van Miltenburg, Buskens, and Raub 2012 for further discussion and additional empirical evidence).

Some lessons for rigorous sociology

Some lessons can be learned from these empirical studies on how to do rigorous sociology. A rather obvious lesson is that a careful and detailed analysis of broad concepts such as "embeddedness" that distinguishes between different levels and mechanisms through which embeddedness affects interactions is not only theoretically but also empirically feasible. Such an analysis yields new insights in the mechanisms underlying embeddedness effects.

A second lesson concerns the interplay between formal theoretical model building and empirical research. Rather simple models of indefinitely repeated games with complete information have been the seminal starting point of much contemporary theoretical model-building in research on cooperation in problematic

social situations. These models focus exclusively on control effects of dyadic embeddedness. In many respects, they have been the starting point, too, of the *Management of Matches*-program. The following step in the development of the program has been to construct models that account not only for dyadic embeddedness but also for network embeddedness, at the beginning once again with an exclusive focus on control effects. Our empirical studies reveal, first, that there is consistent evidence of the control effects of dyadic embeddedness and that effects of network embeddedness should indeed not be neglected. This is a good reason to construct and use more complex theoretical models that account not only for dyadic embeddedness but also for network embeddedness. Second, however, our empirical studies provide considerable evidence that embeddedness at the dyadic as well as at the network level works not only through control effects but also and particularly through learning effects. This led to – and provided justification for – games with incomplete information to be able to account simultaneously for the control effects and the learning effects of embeddedness. Indeed, in various ways, our empirical studies indicate the need for the development of more complex – rather than simpler – theoretical models, in stark contrast to the oft-encountered knee-jerk reaction of empirically-minded sociologists who believe that complex theoretical models are more of a nuisance for generating testable predictions and for empirical research.

Third, our studies highlight the advantages of using alternative and complementary research designs – in our case: survey research, vignette studies, and lab experiments – for repeatedly testing basically the same predictions, employing a strategy sometimes referred to as "triangulation" or "cross validation." One thus hopes to provide cumulative evidence for the robustness of findings (see Levitt and List 2007; Falk and Heckman 2009; Gächter and Thöni 2011; Jackson and Cox 2013 for discussions of this issue). Each research design has its own typical advantages and disadvantages. For example, survey studies cover actual interactions outside the lab and are thus less problematic with respect to external validity. However, they are often less closely related to the underlying theoretical model and require additional and often hard-to-test assumptions to make them suitable as evidence for or against implications from the underlying theoretical model. Also, they often allow for only rather imperfect measurements of core variables, control over the variation in core independent variables is often problematic, and the causal relationship between variables is likewise often doubtful. Lab experiments can be designed so that they closely represent model assumptions, they allow subjects to be given proper incentives, they permit control over variation in core independent variables, and the causal relationship between manipulations and outcome differences is mostly obvious. A disadvantage is that experimental set-ups are often rather artificial. Subjects are typically students who are engaged in abstract interactions, raising concerns about

external validity. Vignette studies are less abstract than lab experiments, while still allowing for control over variation in core independent variables. However, they typically involve hypothetical decisions in hypothetical situations and "incentive compatibility" is problematic.

It therefore makes sense to test predictions repeatedly with different designs, each having specific strengths and shortcomings, to assess the robustness of empirical results and to contribute to distinguishing design effects from robust evidence (see also Buskens 2014). This research strategy complements theoretical pluralism – competing theories – by likewise focusing on "empirical pluralism," namely, complementary research designs.[34] Note, too, that a subsequent step in implementing empirical pluralism would be to repeatedly test the same or at least similar predictions not only by using different designs but also by testing them in different research domains. For example, one could complement tests of embeddedness effects on trust problems in economic exchange with tests of the same effects on trust problems in social exchange.[35]

Finally, our studies fit into the convergence of theoretical models in rigorous sociology that employ rational choice assumptions with mainstream empirical research. For quite some time, it was a weakness of rational choice sociology that close ties with systematic and cumulative empirical research were largely absent (see Green and Shapiro's 1994 influential criticism of rational choice approaches in social science, which mainly targeted the weak links with empirical research). This situation has now changed. For example, Goldthorpe (2000), in his collection of essays, including his programmatic statement "The quantitative analysis of large-scale data sets and rational action theory: For a sociological alliance" (1996), argued forcefully for a better integration of theoretical models into rigorous sociology and empirical research. He pointed out that the problem was not only that rational choice approaches (Goldthorpe: rational action theory – RAT) in sociology had a feeble empirical basis and that rational action theory needed to use empirical regularities revealed by quantitative analysis of data (Goldthorpe: QAD) as *explananda* but, conversely, also that much large-scale survey research needed to employ rational action theory as a tool to provide an *explanans* for those *explananda*. Goldthorpe's plea seems to have encouraged a substantial amount of work by theorists as well as empirical researchers aimed at reducing the gap between rational choice approaches on the one hand and empirical research in sociology on the other (see Blossfeld and

[34] I tend to believe that the biggest impact of the *Management of Matches*-program on research within the ICS and on a number of ICS PhD theses is that it has made the use of complementary designs more common, including more focus on the use of experimental designs.

[35] See De Ruijter (2005) for research on households and behavior of partners in households that tests similar predictions like those on embeddedness effects in a very different domain.

Prein 1998 for an example). Wittek, Snijders, and Nee (2013) is a major handbook taking stock of this development. Empirical work in the *Management of Matches*-program fits in this development. The advantages of employing complementary research designs to test the same predictions also show that it makes sense to extend Goldthorpe's program for setting out an alliance between rational action theory and quantitative analysis of data by conceiving QAD more broadly, including not only survey designs but also, for example, experiments and quasi-experimental designs such as vignette studies.

2.3 Extensions and refinements

I round off my overview of work on game-theoretic models of trust and cooperation in social dilemmas with some remarks on how analyses of embeddedness effects can be and have been extended and further refined.

Cooperation through institutional embeddedness and commitments

Institutions often enhance embeddedness by allowing actors to inform others, thus enhancing opportunities to cooperate conditionally. Modern examples are eBay's feedback forum and similar reputation systems used in the Internet economy. An institution such as the eBay feedback forum allows buyers to evaluate sellers and to collect information on sellers from other buyers. Likewise, sellers can provide and receive feedback on buyers. Similar institutions in medieval trade are the Maghribi traders' coalition (see Greif 2006 for a comprehensive treatment) and the law merchants (Milgrom, North, and Weingast 1990). While they help actors overcome social dilemmas in economic exchange, such institutions cannot be taken for granted, for example, due to incentive problems associated with the provision of (correct) information and feedback. Hence, one of the prominent features of the models provided by Greif and Milgrom et al. is that the institutions are "endogenized" in the sense that it is shown that they are themselves the result of equilibrium behavior in repeated games.

Institutions can help actors overcome social dilemmas in other ways too. Contract law and other institutions often provide opportunities for actors to modify their own (future) incentives themselves, or, as Coleman (1990) put it, to construct their social environment. Actors can do so by incurring *commitments*. Schelling (1960) has shown how incurring commitments ("burn bridges behind one") can strengthen an actor's bargaining position and can thus affect how distribution problems and conflicts are solved (see also Williamson 1985). Similarly, commitments can help actors solve trust and cooperation problems. For example, a seller in the role of the trustee voluntarily provides a guarantee before the Trust Game itself is played. The guarantee modifies the

Thomas Schelling

subsequent incentives of the trustor and trustee in the Trust Game. Commitments such as guarantees can promote trust by reducing the trustee's incentive to abuse trust, by providing compensation for the trustor if trust is abused, or by signaling that the trustee will not (or cannot) abuse trust. Game-theoretic models can be used to specify conditions in which commitments are incurred and induce placing and honoring trust (e.g., Weesie and Raub 1996; Raub 2004). These models make it possible to predict how characteristics of the commitment – such as the costs associated with incurring a commitment, the degree by which the trustee's incentive to abuse trust is reduced, or the size of the compensation for the trustor in the event of abused trust – affect the likelihood of a commitment being incurred as well as the likelihood of trust being placed and honored. In these models, a context that provides opportunities for incurring a commitment is assumed as exogenous. The commitment itself can then be conceived of as a "private institution," voluntarily created by the actors involved in a social dilemma for overcoming the dilemma. One strength of the models is then that the private institution is again not taken for granted but is itself an outcome of equilibrium behavior.

Gideon Keren

Predictions from game-theoretic models of trust and cooperation by voluntarily incurring commitments have been tested experimentally. In fact, my own first hands-on experience with experimental work consisted of experiments with commitments as a mechanism of cooperation for the Prisoner's Dilemma, designed and conducted together with Gideon Keren at Utrecht University in 1989, some years before the Faculty of Social and Behavioral Sciences boasted a well-equipped *Experimental Laboratory for Sociology and Economics* (Raub and Keren 1993). It was fascinating to witness how changing one sentence in a five-page set of instructions for subjects made a sizeable impact on behavior. This was a sentence specifying the conditions under which a subject would lose the commitment, if incurred, and whether a lost commitment would or would not be given to the partner. While leaving this implicit as such in the instructions, the sentence was crucial for determining whether or not incurring a commitment and subsequently cooperating could be regarded as the result of game-theoretic equilibrium behavior and for determining the properties of the equilibrium, such as subgame-perfection. It was likewise fascinating to see that rational models provided clear guidelines on how to proceed with further experimental work, depending

on earlier empirical results. For example, having noticed that commitments were hardly incurred when such behavior was "only" supported by an equilibrium that was not subgame perfect, game theory clearly suggested checking whether things would change when subgame-perfection was "added" – and they did indeed change. Somewhat later, Chris Snijders developed and experimentally tested game-theoretic models for commitments in Trust Games (Snijders 1996).

Institutional embeddedness that provides opportunities for incurring commitments can be a substitute as well as a complement for embeddedness in the sense of ongoing interactions or network embeddedness. Because they can commit, actors can overcome trust problems and other social dilemmas even if ongoing interactions or networks are absent or are insufficient to promote trust, for example, due to large incentives for abusing trust ("golden opportunities"; see Raub 1992 for early work in this direction). Embeddedness and commitment can complement each other in inducing trustfulness and trustworthiness in situations with trust problems. For example, consider a repeated Trust Game with an opportunity for the trustee to incur a commitment at the beginning of each round. Game theory can be used to derive a theorem on conditions for three properties of such a game. First, a subgame-perfect equilibrium exists which implies commitments by the trustee at the beginning of each round together with placing and honoring trust being conditional upon commitments as well as upon trustfulness and trustworthiness in earlier rounds of the repeated game. Second, there are no equilibria that induce trustfulness and trustworthiness without preceding commitments. Third, there are also no equilibria that make placing and honoring trust exclusively conditional upon prior commitments and not also conditional upon trustfulness and trustworthiness in earlier rounds of the game. Given these properties, commitments and conditional cooperation in repeated interactions complement each other and such complementarity is essential for placing and honoring trust by rational actors. Testable predictions for experiments follow from the theorem.

As an aside, game-theoretic models of commitments can be re-interpreted so that they shed light on endogenous preference changes, too (Hegselmann, Raub, and Voss 1986; Raub and Voss 1990). The models show that it can be individually rational for purely self-regarding actors to incur a commitment, thus modifying their own outcomes in a subsequent social dilemma. Imagine now that an actor, rather than being able to manipulate his outcomes, is able to directly manipulate his preferences over outcomes, while the outcomes as such remain the same. It follows that rational and self-regarding actors who are able to choose and modify their own preferences over outcomes in a subsequent social dilemma would be willing to do so.

Rainer Hegselmann

Cooperation in non-repeated social dilemmas

Cooperation in social dilemmas based on dyadic or network embeddedness is a result of reciprocity in repeated interactions: actors condition their behavior in present interactions on what happened in previous interactions, either involving the same partners or other actors. Other mechanisms must obviously be used to explain cooperation in non-repeated interactions. Consider trust in "isolated encounters." Two actors play the Trust Game once and only once. The payoffs from Figure 2 represent monetary incentives for trustor and trustee. Neither the two actors nor other actors can condition their behavior in future interactions on what happens in the Trust Game. Given the standard model of rational behavior and self-regarding preferences, the prediction is no trust and, if trust were to be placed anyway, it would be abused. Similarly, under these assumptions, defection can be predicted in other social dilemmas. Experimental research largely rejects such predictions and shows that defection and opportunism are not ubiquitous in isolated encounters resembling a social dilemma (see, for example, Snijders 1996 and Camerer 2003: Chapter 2.7 for the Trust Game; Berg, Dickhaut, and McCabe 1995 on the Investment Game; Sally 1995 for the Prisoner's Dilemma; Ledyard 1995 for *n*-person dilemmas).

Different approaches are being used that account for such empirical regularities (see Fehr and Schmidt 2006 for an instructive overview). Each of these approaches involves making assumptions about the actors more complex than the standard model in one way or the other. First, one could relax the rationality assumption and employ a *bounded rationality* perspective. For example, one could assume that subjects in experiments are used to repeated interactions in life outside the laboratory. The assumption then is that subjects erroneously apply rules in isolated encounters that are appropriate when interactions are repeated (see, e.g., Binmore 1998 for a sophisticated discussion of such approaches). More generally, Binmore (1998: Chapter 0.4.2) argues that we can expect behavior in experimental games to be consistent with the assumption of selfish game-theoretic rationality only if the game is easy to understand, adequate incentives are provided, and sufficient time is available for trial-and-error learning (see Kreps 1990b for similar arguments).

Second, there are approaches that maintain the rationality assumption but modify the selfishness assumption. These approaches abandon the assumption that actors are purely self-regarding and care only about their own material resources. Rather, it is assumed that at least some actors have *other-regarding preferences*. It is often argued (e.g., Fehr and Gintis 2007) that such preferences are due to socialization processes and internalized social norms and values.

Other-regarding preferences can be modeled in various ways. Some models assume *outcome-based social preferences* (in the sense of Fehr and Schmidt 2006): the

utility function of an actor depends not only on the actor's own material resources but also on the material resources other actors receive. In social psychology, Kelley and Thibaut (1978) pushed this idea much earlier. Influential models (Fehr and Schmidt 1999; Bolton and Ockenfels 2000) assume inequity aversion: an actor's utility increases with his own material resources and if the allocation of material resources becomes more equitable. In the Trust Game with payoffs in terms of money, the trustee may thus honor trust because the utility of honoring trust exceeds the utility of abusing trust. After all, honoring trust yields a monetary payoff R_2 as well as an equitable distribution (R_1, R_2) of monetary payoffs. Conversely, abusing trust yields a higher monetary payoff T_2 but the trustee's utility is reduced due to the inequitable distribution (S_1, T_2). If the disutility from the inequitable distribution is large enough, a utility-maximizing trustee would thus honor trust. In principle, social preferences can be relatively easily accommodated in a game-theoretic framework by employing appropriate assumptions about the actors' utility functions.

Another approach to other-regarding preferences employs the assumption that they are *process-based* rather than outcome-based. The idea is that actors care about *other actors' intentions* (an influential model is due to Rabin 1993). The basic intuition is that if an actor feels that another actor has treated him kindly, he is inclined to return the favor, and that he will wish to hurt the other actor if he has been treated with hostile intentions. In the Trust Game, the trustee may thus honor trust because he perceives the placement of trust by the trustor as kind behavior. Intention-based models of other-regarding preferences can be integrated into a game-theoretic framework but such an integration requires considerable adaptations (see Fehr and Schmidt 2006 for details and references).

The *Management of Matches*-program focused primarily on explaining cooperation through embeddedness effects rather than through non-standard assumptions about actors' preferences. However, Vieth (2009), Aksoy (2013), and Van Miltenburg (2015) accounted for isolated encounters. Vieth studied how commitments can foster cooperation under the assumption of process-based preferences. Van Miltenburg contributed to the literature on sanctioning institutions that developed from Fehr and Gächter (2002), analyzing how sanctioning opportunities after an isolated encounter can foster cooperation among actors when at least some of them have other-regarding preferences. Aksoy proceeded from the idea that actors' behavior in social dilemmas depends on their own preferences as well as their beliefs about other actors' preferences. He analyzed in detail, theoretically as well as experimentally, the role of beliefs about others' preferences.

Assumptions about other-regarding preferences should be used with care (see, e.g., Camerer 2003: 101; Fehr and Schmidt 2006: 618), since almost all behavior can be "explained" by assuming the "right" preferences and adjusting the utility function.

One would first of all prefer parsimonious assumptions about other-regarding preferences, adding as few new parameters as possible to the model. Second, when assumptions on other-regarding preferences are employed, one should aim to use the same set of assumptions to explain behavior in a broad range of different experimental games. Third, one should account not only for well-known empirical regularities but also endeavor to derive and test new predictions. It is therefore important from a methodological perspective that the same set of assumptions about other-regarding preferences is consistent with empirical regularities of behavior in different social dilemmas as well as, for example, in games involving distribution problems such as the Ultimatum Game (Güth, Schmittberger, and Schwarze 1982) or the Dictator Game (Kahneman, Knetsch, and Thaler 1986), and in market games.

Models like those of Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) do indeed try to account for empirical regularities in a broad class of experimental games. They do so by assuming heterogeneity between the actors with respect to their inequity aversion: actors differ with respect to their other-regarding preferences – there are self-regarding actors as well as actors with other-regarding preferences. Moreover, actors are incompletely informed about others' preferences.[36] The interaction between actors who are self-regarding and actors with other-regarding preferences in a setting with incomplete information on other actors' preferences can be a driving force in inducing seemingly purely selfish behavior in settings such as experimental markets and non-selfish behavior in other games, for example, social dilemmas that are isolated encounters (but see Shaked 2006 and Fehr and Schmidt 2005 for an unusually heated debate on whether the Fehr-Schmidt model is indeed successful in these respects).

Strategic network formation and the co-evolution of networks and behavior

A common feature of the game-theoretic models on effects of dyadic and network embeddedness 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 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. Therefore, in addition to analyzing "games on networks" (Goyal 2007: Chapter 3; Jackson 2008: Chapter 9) that can be used to model embeddedness effects, one would like to also analyze "strategic network formation," namely, embeddedness as a result of purposive action.

[36] Therefore, the focus is on games such as the Trust Game with incomplete information.

Until now, there have only been a few game-theoretic models that include such investments and thus endogenize embeddedness by simultaneously modeling investments in and the effects of embeddedness in Trust Games (see Corten 2014 for work on the co-evolution of networks and behavior; Flap 2004 suggested an integrated analysis of investments in and returns on social capital more generally). One class of such models assumes indefinitely repeated Trust Games (Raub, Buskens, and Frey 2013) or finitely repeated Trust Games with incomplete information (Frey, Buskens, and Raub 2015; Frey 2016), 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 Chapter 2.1. The new feature is additional predictions on investments in network embeddedness.

The core result, robust when assuming either indefinitely repeated Trust Games or finitely repeated Trust Games with incomplete information, 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 small trust problems that can be mitigated through dyadic embeddedness alone. They 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 on top of dyadic embeddedness make a difference in 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 of conditional trust compared to no trust exceed the costs of the investments. Frey, Buskens, and Corten (2016) offer preliminary experimental work testing predictions from such models.

Rense Corten

Vincenz Frey

These theoretical results apply for trustors as well as trustees. All actors are better off when trust is placed and honored compared to when no trust is 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, equilibrium behavior may consist of the trustee himself investing in setting up an information network for trustors, credibly committing to honoring trust in this way. For example, not only buyers but also sellers in online markets have an interest in the availability of a reputation system.

Neuroscience in the social sciences

Since around 2000, a sizeable literature has emerged at the interface of the social sciences and neuroscience. It includes neuroeconomics, often driven by economists who became interested in neuroscience applications in their own discipline and thus started to collaborate with neuroscientists (see Glimcher and Fehr 2014). A somewhat complementary research line is social neuroscience, often driven by neuroscientists who became interested in applications of their own approach in new fields. Specifically, they became interested in neuroscience applications in research on social interactions, broadly conceived (see Decety and Cacioppo 2015). Sociology seems to be catching up and a handbook of neurosociology has meanwhile become available (Franks and Turner 2013). This work is interesting from the perspective of rigorous sociology and, more specifically, interesting for research on social dilemmas. After all, neuroscience can be expected to contribute to a better understanding of the regularities underlying human behavior. Also, game theory as a theory of human behavior has provided new insights for research on social dilemmas, but we have seen that predictions are not always supported empirically and that some empirical regularities are hard to reconcile with game-theoretic models. It therefore makes sense to explore the potential for joining forces.

We have meanwhile conducted some work in this field (for example, Buskens et al. 2016), using experimental designs employing drug administration and focusing on testosterone effects in social interactions. This research has two characteristics. First, we try to systematically generate predictions of testosterone effects from underlying theoretical assumptions and test those predictions rather than search for testosterone effects in a more or less inductive fashion. Second, we focus on the interplay of social macro-conditions and testosterone levels, thus addressing interaction effects of macro conditions and testosterone levels, rather than focusing exclusively on the main effects of "sociological" and "(neuro)biological" variables. While the main effects of testosterone levels on behavior in social interactions are certainly interesting, including from a sociological perspective, it seems that the case for interdisciplinary work in this field becomes stronger when it focuses on such interaction effects and when such effects are found.

Buskens et al. (2016) studied testosterone effects on behavior in trust problems, employing a variant of the Investment Game. The study is based on two ideas. First, one of the core effects of testosterone seems to be that it impairs cognitive empathy (Theory of Mind, mind reading, cognitive perspective taking; see, e.g., Singer and Tusche 2014). This should influence behavior in strategic situations such as the Trust Game or the Investment Game since the behavior of an actor in such situations is likely to depend on the actor's anticipation of the other actor's behavior and, thus, to depend on cognitive empathy. Second, we employ that, roughly speaking, cognitive empathy is a considerably more demanding task for the trustor in the repeated game than in the non-repeated game. This is because her own behavior in the current round of the repeated game can have repercussions for behavior in multiple future rounds. Moreover, the trustor's cognitive empathy in the repeated game requires her not only to anticipate how her own behavior in the current round may affect the behavior of the trustee in the current and future rounds; it now also implies an additional step of strategic reasoning, namely, that the trustor also anticipates whether and how the trustee, in turn, might anticipate on the effects of trustee behavior for subsequent trustor behavior. One would thus expect an interaction of testosterone with the social condition "non-repeated versus repeated game" such that testosterone will reduce the difference in trustor behavior between the non-repeated and the repeated game. The experiment indeed provides some preliminary evidence in favor of such a prediction, well in line with related studies on testosterone effects in social interactions (e.g., Van Honk et al. 2011; Carré et al. 2015).

Rational models for running an academic institution

Those running an academic institution are often involved in trust and cooperation problems themselves with other members of these institutions. Also, they face the problem of providing those who are running subgroups within the institution with means for solving trust and cooperation problems in these subgroups. Research shows that incurring commitments can be a mechanism for solving trust and cooperation problems. This implies that those running an academic institution can have good reasons to commit themselves. Such commitments affect the future behavior of those who commit themselves and, particularly, the future behavior of those with whom they interact, provided that the commitments possess certain characteristics. For example, in trust problems, a trustee's commitment must convince the trustor that the trustee is trustworthy.

Consider the Executive Board of a Faculty in the role of a trustee vis-à-vis, say, the Heads of Departments as trustors. First, rather than being ambiguous with

respect to own future behavior, it makes sense for the Board to be clear and specific about its policies. One effect of being clear and specific rather than ambiguous is that deviations by the Board from clear and specific policies are visible, thus negatively affecting the Board's reputation, making it less attractive for the Board to deviate, and inducing the Heads of Departments to cooperate. In addition, another rationale emerges for striking a balance between a large lump-sum budget for Departments and reserving large portions of the budget for separate projects on the Faculty level, so that the lump-sum for Departments is relatively large. After all, a policy of this kind commits the Board by decreasing its discretion. At the same time, the Heads of Departments receive additional means to solve trust and cooperation problems within their respective Departments. Of course, such policies presuppose that both the Executive Board and the Heads of Departments understand how commitments can contribute to solving trust and cooperation problems in social dilemmas. One direct implication is the importance of selection procedures and training facilities that foster such competencies.

The research on cooperation in non-repeated dilemmas shows that when an actor's preferences are clearly signaled, other actors' behavior can be affected. In trust problems, clear signals that a trustee has other-regarding rather than purely self-regarding preferences can induce a trustor to place trust. Also, given process-based preferences, reciprocity is a powerful driver of behavior not only in repeated but also in non-repeated interactions. This highlights that those running academic institutions are well advised to be cooperative themselves, including in the sense of taking the initiative in complex situations (in Dutch: *in de strijd voorop gaan* and *knopen doorhakken*), since this will often induce others to follow.

The final conclusion is rather obvious but nevertheless worth mentioning. Empirical research in the social and behavioral sciences – survey research as well as experimental studies – presupposes an adequate and costly technical infrastructure. Moreover, such an infrastructure is not only necessary for empirical research as such but likewise allows for progress in theory formation since theoretical progress depends on a close relationship between theory and empirical research. It is easy, though not off the mark, to conclude that this requires adequate budgets not only for the "hard" sciences but also for the social and behavioral sciences (and without doubt for the humanities, too). The less easy but relevant conclusion is that given the available scarce budget, investments in research infrastructure unavoidably imply that fewer resources are available for other useful initiatives: there will be less office space than in an academic paradise and a better research infrastructure will mean fewer PhD students and postdocs. Conversely, it is likewise necessary to use costly research infrastructure efficiently. That means that dedicated labs for individual researchers or chairs are not an obvious way to go.

3. Conclusions: What has been achieved?

Summarizing, what has been achieved in rigorous sociology in general and more specifically in research using game-theoretic models of social dilemmas over roughly the past four decades? It seems to me quite clear that progress has been made, in general, in micro-macro modeling. Rigorous sociology as a whole has improved with respect to theoretical and methodological rigor, the concern for testable predictions and the convergence of theory formation, empirical research, and statistical modeling.

Game-theoretic models of social dilemmas have yielded a better understanding of mechanisms that can drive trust and cooperation. This includes a better understanding of the conditions whereby even purely self-regarding rational actors can cooperate as well as a better understanding of the limits of such cooperation. In sociology, Parsons and Durkheim correctly pointed out that cooperation in situations resembling social dilemmas posed challenging explanatory problems for theoretical approaches using assumptions about goal-directed and incentive-driven behavior. The alternative that developed in the Parsons-tradition in sociology focused on unconditional norms and values, learned and internalized in a process of socialization, as a solution to the Hobbesian problem of social order. By the 1970s, it had become clear that this "normative solution" to the Hobbesian problem of social order had its own serious problem such as overestimating the impact of such norms and circumventing an explanation of how those norms emerge and are being maintained and enforced (Wrong 1961; Coleman 1964b; Elias 1970; Vanberg 1978). The rational choice models of trust and cooperation, including game-theoretic models, that have evolved since then have made progress in answering Parsons' challenge by specifying the rationale underlying the challenge and the conditions under which purposive actors can cooperate.

It should be stressed that progress has also been made with respect to empirical research and empirical insights. Abstract game-theoretic models do allow for generating empirically testable predictions and such predictions have been subjected to systematic empirical research. This fits in with a more general pattern that seems typical of the development of rigorous sociology, namely, the convergence of theory formation, including theory using micro-level assumptions about purposive action ("RAT" in Goldthorpe's sense) and empirical research (Goldthorpe's "QAD"), arguably a major advancement in the discipline in recent decades. What is more, it has been feasible to test game-theoretic predictions of trust and cooperation using complementary research designs, thus establishing robust evidence and in the process broadening Goldthorpe's QAD by including various experimental designs in addition to survey studies.

A considerable amount of empirical evidence for embeddedness effects has meanwhile become available. When research designs are employed that disentangle

different kinds of embeddedness effects and the mechanisms through which embeddedness works, hypotheses based on game-theoretic models often succeed in predicting the signs of coefficients (see, e.g., Grofman 1993 and Green and Shapiro 1994 for related discussion of the merits and problems of qualitative predictions on changes "at the margin" using comparative statics versus quantitative point predictions from rational choice models). Nevertheless, there is room to improve the predictions of game-theoretic models on behavior in social dilemmas. The overall impression is that assuming game-theoretic rationality as well as self-regarding actors cannot account for much of the cooperation in social dilemmas that are isolated encounters, while such assumptions often predict "too much" cooperation in repeated social dilemmas (see Bolton and Ockenfels 2009 for a similar point in the context of research on reputation systems in the Internet economy).[37] Developing game-theoretic models on the interplay of social embeddedness and other-regarding preferences may be useful in this respect (see, e.g., Gintis 2000: Chapter 11 for related arguments).

Given the empirical evidence, different alternatives are feasible for theory formation. In terms of Coleman's diagram and following Coleman's (1987b) own strategy, one can try to refine and improve transformation rules and bridge assumptions that link macro- and micro-levels of analysis. Alternatively, one can try to adapt micro-level assumptions about regularities of behavior. I have sketched that this could involve adapting rationality assumptions as well as assumptions about actors' preferences. Behavioral and experimental game theory (see Camerer 2003 for a textbook) provides examples of such work, as well as work on learning in games and work in evolutionary game theory (see Fudenberg and Levine 1998 and Gintis 2000 for textbooks). Within sociology, micro-models such as Esser's (1996) framing model and Lindenberg's (2001) theory of social rationality have been proposed as alternatives for standard rational choice models.

When one wishes to come up with micro-models superior to the standard game-theoretic 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 give better accounts of the overall patterns of empirical evidence for social dilemmas rather than exclusively accounting for some specific empirical "anomaly" relative to game-theoretic predictions. Furthermore, one would need an alternative that, like game theory, models a core feature, namely interdependence between actors and comes up with assumptions about how interdependence affects

[37] It is clear, though, that *if* empirical findings differ from model predictions, the deviations *have to be* in these directions, since the model predictions are, in principle, no cooperation at all in isolated encounters and cooperation throughout in repeated interactions (assuming that appropriate conditions for repeated interactions are fulfilled).

behavior, including assumptions about how actors anticipate others' behavior. Available alternatives 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 worthwhile to bet on theoretical pluralism and foster the development of theoretical alternatives rather than to dismiss alternatives prematurely.

Structural individualism was mentioned at the beginning of this lecture as a research program in European sociology that provided the foundations for much of the work that I sketched, as well as guidelines on how to make incremental and cumulative progress in such work. It has been a pleasure to experience that these foundations have proven to be solid and the guidelines to be fruitful.

4. Coda

Acknowledgments at the end of a lecture like this are a problem: too many colleagues and friends to mention, while time and space are limited.

The Executive Board of Utrecht University and the Executive Board of the Faculty of Social and Behavioral Sciences had sufficient confidence and trust to appoint me as Chair of Theoretical Sociology, including the opportunities that come with such an appointment. The Executive Board of our university even mustered sufficient confidence and trust to appoint me as Dean of Social and Behavioral Sciences. *Pars pro toto*, I would like to mention Willem Hendrik Gispen, our former Rector, and Jan Rispens, former Dean of Social and Behavioral Sciences. Both provided support when I was responsible for the Department of Sociology. And Bert van der Zwaan, our current Rector, with whom collaboration has been a pleasure and has often been "lively" during my Deanship. I hope I have lived up to the expectations associated with my appointments.

NWO, *The Netherlands Organisation for Scientific Research*, has been a supportive institution, particularly through generous funding for the *Management of Matches*-program.

During my academic career, I have benefited from my mentors, co-authors, and my PhD students. References to a few of them in my lecture, including the snapshots, were intended as acknowledgments, too. I learned from all of them. I have been particularly lucky, from the start of my academic career as an undergraduate student and for a long period throughout that career, with all my mentors: Rainer Hegselmann, Carl-Friedrich Gethmann, Hartmut Esser, Hans Hummell, Günter Büschges, and Reinhard Wippler. As co-authors, I mention *pars pro toto* Thomas Voss, Jeroen Weesie, and Vincent Buskens. With respect to PhD students, one of the many things I learned from Reinhard Wippler was to see to it that they develop their own research agenda. The likelihood of "I shall not wholly die" (Horatius, Ode 3.30) because one has spread one's scientific genes through one's PhD students increases when students have and make use of this freedom, rather than when they have to carry out projects that have been pre-designed in great detail. I'm delighted to have had PhD students who did develop their own research agenda, many of them now pursuing their own academic careers.

I enjoyed working with many other colleagues. The Department of Sociology and the ICS have been outstanding and supportive academic institutions – my intellectual home and community since 1989, when I joined the Utrecht Faculty for the first time. In fact, it was my intellectual home quite a bit earlier, when I started to make contact with Reinhard Wippler, Frits Tazelaar, and Jeroen Weesie in the early 1980s. Quite some of my remarks on how to run an academic institution

are a result of "reverse engineering": observing what worked in the Department of Sociology and in the ICS, that was the context of discovery, to use a perspective from analytical philosophy of science. Then came the context of justification: searching for the rationale behind why it worked. When I had a leading role in the Department of Sociology and the ICS, I shared responsibilities in Utrecht in various periods with colleagues such as Harry Ganzeboom, Louk Hagendoorn, Henk Flap, and Tanja van der Lippe, during a crucial period benefiting from first-class administrative support provided by our Managing Director Agnes Barentsen. In the ICS, I shared responsibilities with colleagues in Groningen such as Frans Stokman, Siegwart Lindenberg, Tom Snijders, and Rafael Wittek. It's a delight that the Department and the ICS have remained in good shape in recent years, due very much to the way in which colleagues such as Tanja van der Lippe, Vincent Buskens, and Frank van Tubergen are operating. It's a delight, too, that the new funding from the NWO Gravitation Program for SCOOP – *Sustainable Cooperation* ensures a promising future, including bright prospects for new research closely related to the topics of this lecture.

As Dean, I enjoyed collaborating with colleagues at Utrecht University in the same role, particularly with Wiljan van den Akker and Keimpe Algra as well as Henk Kummeling and Annetje Ottow. I enjoyed collaborating with Chantal Kemner and Bas van Bavel, who are running our Strategic Themes *Dynamics of Youth* and *Institutions for Open Societies*. I enjoyed collaborating with colleagues in our *Bestuursgebouw* such as Joop Kessels, Leon van de Zande, and Esther Stiekema – their expertise and dedication to high standards have been very supportive.

While serving as Dean, I had the privilege that I could share responsibility with our Vice-Deans Theo Wubbels, Susan te Pas, and Marcel van Aken. Throughout, we had outstanding support from Henk van Rinsum, Martine Verbeek, and Mariska Phielix as Secretaries to the Board. It was a pleasure to work with all of them and also to work with our support staff, with everybody else who is responsible for the ups and downs of various parts of our Faculty, with the Faculty Council (our *Faculteitsraad*), all other colleagues, and with our students.

Last but not least, I would like to mention two more colleagues, also for more personal reasons. Madelon Pieper is one of them, my secretary (in fancier language: my "personal assistant") who managed my appointments calendar, among many, many other things. Every now and then, I received compliments for being well organized. That has been largely due to Madelon.

And finally, I have not yet mentioned one member of the Executive Board of our Faculty, namely our Director Frank Jan van Dijk. I wanted him to be the very last person who I mention by name – the last one at these occasions is always special, and Frank Jan is certainly special. Running the Faculty for more than five years was

hard work. Being able to collaborate with an outstanding Director like Frank Jan made it much easier. Often enough, though, running the Faculty was also rewarding and sometimes even a pleasure. Frank Jan contributed a lot to that, too. Dear Frank Jan, I was privileged indeed to work with you and rely on you, on your expertise, your dedication, and your good judgment. I'll miss our collaboration.

While I am looking much forward to my future at our university with more time for research than in the previous period, let me thank all those with whom I have worked together in all my years at Utrecht University, since 1989, with some breaks. Thanks for all your cooperation. I hope that I have been able to reciprocate at times.

Ik heb gezegd.

References

Akerlof, George A. (1970) "The market for 'lemons': Quality uncertainty and the market mechanism." *Quarterly Journal of Economics* 89: 488–500.

Aksoy, Ozan (2013) *Essays on Social Preferences and Beliefs in Non-Embedded Social Dilemmas*. PhD thesis, Utrecht University.

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

Auspurg, Katrin, and Thomas Hinz (2015) *Factorial Survey Experiments*. Los Angeles, CA: Sage.

Axelrod, Robert (1984) *The Evolution of Cooperation*. New York: Basic Books.

Bacharach, Michael, and Diego Gambetta (2001) "Trust in signs." Pp. 148–184 in *Trust in Society*, edited by Karen S. Cook. New York: Russell Sage.

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

Baron, James N., and David M. Kreps (1999) *Strategic Human Resources. Frameworks for General Managers*. New York: Wiley.

Batenburg, Ronald S., Werner Raub, and Chris Snijders (2003) "Contacts and contracts: Temporal embeddedness and the contractual behavior of firms." *Research in the Sociology of Organizations* 20: 135–188.

Becker, Gary S. (1976) *The Economic Approach to Human Behavior*. Chicago, IL: University of Chicago Press.

Bendor, Jonathan, and Dilip Mookherjee (1987) "Institutional structure and the logic of ongoing collective action." *American Political Science Review* 81: 129–154.

Berg, Joyce, John Dickhaut, and Kevin McCabe (1995) "Trust, reciprocity, and social history." *Games and Economic Behavior* 10: 122–142.

Bernheim, B. Douglas, and Erik Madsen (2017) "Price cutting and business stealing in imperfect cartels." *American Economic Review* 107: 387–424.

Billari, Francesco C. (2015) "Integrating macro- and micro-level approaches in the explanation of population change." *Population Studies* 69(S1): S11–S20.

Binmore, Ken (1998) *Game Theory and the Social Contract, Volume 2: Just Playing*. Cambridge, MA: MIT Press.

Blau, Peter M. (1964) *Exchange and Power in Social Life*. New York: Wiley.

Blossfeld, Hans-Peter, and Gerald Prein, eds. (1998) *Rational Choice Theory and Large-Scale Data Analysis*. Boulder, CO: Westview.

Blumberg, Boris F. (1997) *Das Management von Technologiekooperationen [The Management of R&D Alliances]*. PhD thesis, Utrecht University.

Bohnen, Alfred (1975) *Individualismus und Gesellschaftstheorie [Individualism and Social Theory]*. Tübingen: Mohr.

Bolton, Gary E., and Axel Ockenfels (2000) "ERC: A theory of equity, reciprocity, and competition." *American Economic Review* 90: 166–193.

Bolton, Gary E., and Axel Ockenfels (2009) "The Limits of Trust in Economic Transactions. Investigations of Perfect Reputation Systems." Pp. 15–36 in *eTrust: Forming Relationships in the Online World*, edited by Karen S. Cook, Chris Snijders, Vincent Buskens, and Coye Cheshire. New York: Russell Sage.

Boudon, Raymond (1974) *Education, Opportunity, and Social Inequality*. New York: Wiley.

Boudon, Raymond (1977) *Effets pervers et ordre social*. Paris: PUF. English translation: *The Unintended Consequences of Social Action*. London: Macmillan 1982.

Boudon, Raymond (1979) *La logique du social*. Paris: Hachette. English translation: *The Logic of Social Action*. London: Routledge 1981.

Burt, Ronald S., and Marc Knez (1995) "Kinds of third-party effects on trust." *Rationality and Society* 7: 255–292.

Buskens, Vincent (2002) *Social Networks and Trust*. Boston, MA: Kluwer.

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

Buskens, Vincent (2014) *Coöperatie in context: Experimentele sociologie 2.0 [Cooperation in Context: Experimental Sociology 2.0]*. Inaugural lecture, Utrecht University.

Buskens, Vincent, Vincenz Frey, and Werner Raub (2017) "Trust games: Game-theoretic approaches to embedded trust." Forthcoming in *The Oxford Handbook of Social and Political Trust*, edited by Eric M. Uslaner. Oxford: Oxford University Press.

Buskens, Vincent, and Werner Raub (2002) "Embedded trust: Control and learning." *Advances in Group Processes* 19: 167–202.

Buskens, Vincent, and Werner Raub (2013) "Rational Choice research on social dilemmas." Pp. 113–150 in *Handbook of Rational Choice Social Research*, edited by Rafael Wittek, Tom A.B. Snijders, and Victor Nee. Stanford, CA: Stanford University Press.

Buskens, Vincent, Werner Raub, Nynke van Miltenburg, Estrella R. Montoya, and Jack van Honk (2016) "Testosterone administration moderates effect of social environment on trust in women depending on second-to-fourth digit ratio." *Nature Scientific Reports* 6: 27655.

Buskens, Vincent, Werner Raub, and Joris van der Veer (2010) "Trust in triads: an experimental study." *Social Networks* 32: 301–312.

Buskens, Vincent, Werner Raub, and Jeroen Weesie (2000) "Networks and contracting in information technology transactions." Pp. 77–81 in *The Management of Durable Relations: Theoretical and Empirical Models for Organizations and Households*, edited by Werner Raub and Jeroen Weesie. Amsterdam: Thela Thesis.

Buskens, Vincent, and Jeroen Weesie (2000) "An experiment on the effects of embeddedness in trust situations: Buying a used car." *Rationality and Society* 12: 227–253.

Calvert, Randall (1995) "Rational actors, equilibrium, and social institutions." Pp. 57–94 in *Explaining Social Institutions*, edited by Jack Knight and Itai Sened. Ann Arbor, MI: University of Michigan Press.

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

Camerer, Colin F., and Keith Weigelt (1988) "Experimental tests of a sequential equilibrium reputation model." *Econometrica* 56: 1–36.

Carré, Justin M., Triana L. Ortiz, Brandy Labine, Benjamin J.P. Moreau, Essi Viding, Craig S. Neumann, and Bernard Goldfarb (2015) "Digit ratio (2D:4D) and psychopathic traits moderate the effect of exogenous testosterone on socio-cognitive processes in men." *Psychoneuroendocrinology* 62: 319–326.

Coleman, James S. (1964a) *Introduction to Mathematical Sociology*. New York: Free Press.

Coleman, James S. (1964b) "Collective decisions." *Sociological Inquiry* 34(Spring): 166–181.

Coleman, James S. (1986a) "Micro foundations and macrosocial theory." Pp. 345–363 in *Approaches to Social Theory*, edited by Siegwart Lindenberg, James S. Coleman, and Stefan Nowak. New York: Russell Sage.

Coleman, James S. (1986b) "Social theory, social research, and a theory of action." *American Journal of Sociology* 91: 1309–1335.

Coleman, James S. (1987a) "Microfoundations and macrosocial behavior." Pp. 153–173 in *The Micro-Macro Link*, edited by Jeffrey C. Alexander, Bernhard Giesen, Richard Münch, and Neil J. Smelser. Berkeley, CA: University of California Press.

Coleman, James S. (1987b) "Psychological structure and social structure in economic models." Pp. 181–185 in *Rational Choice. The Contrast between Economics and Psychology*, edited by Robin M. Hogarth and Melvin W. Reder. Chicago, IL: University of Chicago Press.

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

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

Corten, Rense (2014) *Computational Approaches to Studying the Co-evolution of Networks and Behavior in Social Dilemmas*. Chichester: Wiley.

Corten, Rense, Stephanie Rosenkranz, Vincent Buskens, and Karen S. Cook (2016) "Reputation effects in social networks do not promote cooperation: An experimental test of the Raub & Weesie model." *PLoS ONE* 11(7): e0155703.

Craswell, Richard (1993) "On the uses of 'trust'." *Journal of Law and Economics* 36: 487–500.

Darley, John M., and Bibb Latané (1968) "Bystander intervention in emergencies: Diffusion of responsibility." *Journal of Personality and Social Psychology* 8: 377–383.

Dasgupta, Partha (1988) "Trust as a commodity." Pp. 49–72 in *Trust: Making and Breaking Cooperative Relations*, edited by Diego Gambetta. Oxford: Blackwell.

Decety, Jean, and John T. Cacioppo, eds. (2015) *The Oxford Handbook of Social Neuroscience*. Oxford: Oxford University Press.
Diekmann, Andreas (1985) "Volunteer's dilemma." *Journal of Conflict Resolution* 29: 605–610.
Diekmann, Andreas (2004) "The power of reciprocity." *Journal of Conflict Resolution* 48: 487–505.
Diekmann, Andreas, and Thomas Voss (2004) "Die Theorie rationalen Handelns. Stand und Perspektiven [The theory of rational action. State of the art and perspectives]." Pp. 13–29 in *Rational Choice Theorie in den Sozialwissenschaften. Probleme und Perspektiven. Rolf Ziegler zu Ehren*, edited by Andreas Diekmann and Thomas Voss. München: Oldenbourg.
Durkheim, Emile ([1893] 1973) *De la Division du Travail Social*. 9th ed., Paris: PUF.
Durkheim, Emile ([1895] 1982) *The Rules of Sociological Method*. Houndmills: Macmillan.
Elias, Norbert (1970) *Was ist Soziologie?* München: Juventa. English translation: *What Is Sociology?* New York: Columbia University Press 1984.
Esser, Hartmut (1996) "Die Definition der Situation [The definition of the situation]." *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 48: 1–34.
Falk, Armin, and James J. Heckman (2009) "Lab experiments are a major source of knowledge in the social sciences." *Science* 326: 535–538.
Fehr, Ernst, and Simon Gächter (2002) "Altruistic punishment in humans." *Nature* 415(6868): 137–140.
Fehr, Ernst, and Herbert Gintis (2007) "Human motivation and social cooperation: Experimental and analytical foundations." *Annual Review of Sociology* 33: 43–64.
Fehr, Ernst, and Klaus M. Schmidt (1999) "A theory of fairness, competition, and cooperation." *Quarterly Journal of Economics* 114: 817–868.
Fehr, Ernst, and Klaus M. Schmidt (2005) *The Rhetoric of Inequity Aversion – A Reply*. Working paper.
Fehr, Ernst, and Klaus M. Schmidt (2006) "The economics of fairness, reciprocity and altruism – experimental evidence and new theories." Pp. 615–691 in *Handbook of the Economics of Giving, Altruism and Reciprocity*, edited by Serge-Christophe Kolm and Jean Mercier Ythier. Amsterdam: Elsevier.
Flache, Andreas (2002) "The rational weakness of strong ties." *Journal of Mathematical Sociology* 26: 189–216.
Flache, Andreas, and Michael W. Macy (2002) "Stochastic collusion and the power law of learning: A general reinforcement learning model of cooperation." *Journal of Conflict Resolution* 46: 629–653.
Flap, Henk (2004) "Creation and returns of social capital." Pp. 3–23 in *Creation and Returns of Social Capital*, edited by Henk Flap and Beate Völker. London: Routledge.
Franks, David, and Jonathan H. Turner, eds. (2013) *Handbook of Neurosociology*. Heidelberg: Springer.
Frey, Vincenz (2016) *Network Formation and Trust*. PhD thesis, Utrecht University.
Frey, Vincenz, Vincent Buskens, and Werner Raub (2015) "Embedding trust: A game-theoretic model for investments in and returns on network embeddedness." *Journal of Mathematical Sociology* 39: 39–72.
Frey, Vincenz, Vincent Buskens, and Rense Corten (2016) *Investments in and Returns on Embeddedness: An Experiment with Trust Games*. Working Paper, Utrecht University.
Friedman, James W. (1971) "A non-cooperative equilibrium for supergames." *Review of Economic Studies* 38: 1–12.
Friedman, James W. (1990) *Game Theory with Applications to Economics*. 2nd ed., New York: Oxford University Press.
Fudenberg, Drew, and David K. Levine (1998) *The Theory of Learning in Games*. Cambridge, MA: MIT Press.
Fudenberg, Drew, and Eric Maskin (1986) "The folk theorem in repeated games with discounting or with incomplete information." *Econometrica* 54: 533–554.
Gächter, Simon, and Christian Thöni (2011) "Micromotives, microstructure, and macrobehavior: The case of voluntary cooperation." *Journal of Mathematical Sociology* 35: 26–65.
Gambetta, Diego (1993) *The Sicilian Mafia. The Business of Private Protection*. Cambridge, MA: Harvard University Press.
Gambetta, Diego, and Steffen Hertog (2016) *Engineers of Jihad*. Princeton, NJ: Princeton University Press.
Gintis, Herbert (2000) *Game Theory Evolving*. Princeton, NJ: Princeton University Press.
Glimcher, Paul W., and Ernst Fehr, eds. (2014) *Neuroeconomics*. 2nd ed., Amsterdam: Elsevier.
Goldthorpe, John H. (1996) "The quantitative analysis of large-scale data sets and rational action theory: for a sociological alliance." *European Sociological Review* 12: 109–126 (reprint pp. 94–114 in Goldthorpe 2000).

Goldthorpe, John H. (2000) *On Sociology. Numbers, Narratives, and the Integration of Research and Theory.* Oxford: Oxford University Press.
Goldthorpe, John H. (2016) *Sociology as a Population Science*. Cambridge: Cambridge University Press.
Gouldner, Alvin W. (1960) "The norm of reciprocity." *American Sociological Review* 25: 161–178.
Goyal, Sanjeev (2007) *Connections. An Introduction to the Economics of Networks.* Princeton, NJ: Princeton University Press.
Granovetter, Mark S. (1985) "Economic action and social structure: The problem of embeddedness." *American Journal of Sociology* 91: 481–510.
Green, Donald P., and Ian Shapiro (1994) *Pathologies of Rational Choice Theory*. New Haven, CN: Yale University Press.
Green, Edward J., and Robert H. Porter (1984) "Noncooperative collusion under imperfect price information." *Econometrica* 52: 87–100.
Greif, Avner (2006) *Institutions and the Path to the Modern Economy: Lessons from Medieval Trade.* Cambridge: Cambridge University Press.
Grofman, Bernard (1993) "Is turnout the paradox that ate rational choice theory?" Pp. 93–103 in *Information, Participation, and Choice*, edited by Bernard Grofman. Ann Arbor, MI: University of Michigan Press.
Güth, Werner, Rolf Schmittberger, and Bernd Schwarze (1982) "An experimental analysis of ultimatum bargaining." *Journal of Economic Behavior and Organization* 3: 367–388.
Hardin, Garrett (1968) "The tragedy of the commons." *Science* 162: 1243–1248.
Hardin, Russel (1982) "Exchange theory on strategic bases." *Social Science Information* 21: 251–272.
Harsanyi, John C. (1967/68) "Games with incomplete information played by 'Bayesian' players I-III." *Management Science* 14: 159–182, 320–334, 486–502.
Harsanyi, John C. (1976) *Essays on Ethics, Social Behavior, and Scientific Explanation*. Dordrecht: Reidel.
Harsanyi, John C. (1977) *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge: Cambridge University Press.
Hedström, Peter (2005) *Dissecting the Social. On the Principles of Analytical Sociology*. Cambridge: Cambridge University Press.
Hedström, Peter, and Peter Bearman, eds. (2009) *The Oxford Handbook of Analytical Sociology*. Oxford: Oxford University Press.
Hedström, Peter, and Lars Udehn (2009) "Analytical sociology and theories of the middle range." Pp. 25–47 in *The Oxford Handbook of Analytical Sociology*, edited by Peter Hedström and Peter Bearman Oxford: Oxford University Press.
Hegselmann, Rainer, Werner Raub, and Thomas Voss (1986) "Zur Entstehung der Moral aus natürlichen Neigungen [From natural to moral preferences]." *Analyse & Kritik* 8: 150–177.
Hirschman, Albert O. (1970) *Exit, Voice, and Loyalty. Responses to Decline in Firms, Organizations, and States*. Cambridge, MA: Harvard University Press.
Hirshleifer, David, and Eric Rasmusen (1989) "Cooperation in a repeated Prisoner's Dilemma with ostracism." *Journal of Economic Behavior and Organization* 12: 87–106.
Hobbes, Thomas ([1651] 1991) *Leviathan*. Cambridge: Cambridge University Press.
Homans, George C. (1958) "Social behavior as exchange." *American Journal of Sociology* 63: 597–606.
Van Honk, Jack, Dennis J. Schutter, Peter A. Bos, Anne-Wil Kruijt, Eef G. Lentjes, and Simon Baron-Cohen (2011) "Testosterone administration impairs cognitive empathy in women depending on second-to-fourth digit ratio." *Proceedings of the National Academy of Sciences U.S.A.* 108: 3448–3452.
Hummell, Hans J. (1973) "Methodologischer Individualismus, Struktureffekte und Systemkonsequenzen [Methodological individualism, structural effects, and system level consequences]." Pp. 61–134 in *Probleme der Erklärung sozialer Prozesse II: Soziales Verhalten und soziale Systeme*, edited by Karl-Dieter Opp and Hans J. Hummell. Frankfurt a.M.: Athenäum.
Hummell, Hans J., and Karl-Dieter Opp (1971). *Die Reduzierbarkeit von Soziologie auf Psychologie. Eine These, ihr Test und ihre theoretische Bedeutung [The Reducibility of Sociology to Psychology. A Thesis, Its Test, and Its Theoretical Relevance]*. Braunschweig: Vieweg.
Jackson, Matthew O. (2008) *Social and Economic Networks*. Princeton, NJ: Princeton University Press.

Jackson, Michelle, and David R. Cox (2013) "The principles of experimental design and their application in sociology." *Annual Review of Sociology* 39: 27–49.
Kahneman, Daniel, Jack L. Knetsch, and Richard Thaler (1986) "Fairness as a constraint on profit seeking: Entitlements in the market." *American Economic Review* 76: 728–741.
Kahneman, Daniel, and Amos Tversky, eds. (2000) *Choices, Values and Frames*. Cambridge: Cambridge University Press.
Kelley, Harold H., and John W. Thibaut (1978) *Interpersonal Relations. A Theory of Interdependence*. New York: Wiley.
Kreps, David M. (1990a) "Corporate culture and economic theory." Pp. 90–143 in *Perspectives on Positive Political Economy*, edited by James E. Alt and Kenneth A. Shepsle. Cambridge: Cambridge University Press.
Kreps, David M. (1990b) *Game Theory and Economic Modelling*. Oxford: Clarendon Press.
Kreps, David M., and Robert Wilson (1982) "Sequential equilibria," *Econometrica* 50: 863–894.
Ledyard, John O. (1995) "Public goods: A survey of experimental research." Pp. 111–194 in *The Handbook of Experimental Economics*, edited by John. H. Kagel and Alvin E. Roth. Princeton, NJ: Princeton University Press.
Levitt, Steven D., and John A. List (2007) "What do laboratory experiments measuring social preference reveal about the real world?" *Journal of Economic Perspectives* 21: 153–174.
Lieberson, Stanley, and Joel Horwich (2008) "Implication analysis: A pragmatic proposal for linking theory and data in the social sciences." *Sociological Methodology* 38: 1–50.
Lindenberg, Siegwart (1977) "Individuelle Effekte, kollektive Phänomene und das Problem der Transformation [Individual effects, collective phenomena, and the problem of transformation]." Pp. 46–84 in *Probleme der Erklärung sozialen Verhaltens*, edited by Klaus Eichner and Werner Habermehl. Meisenheim a.G.: Hain.
Lindenberg, Siegwart (1981) "Erklärung als Modellbau [Explanation as model building]." Pp. 20–35 in *Soziologie in der Gesellschaft*, edited by Werner Schulte. Bremen: Zentraldruckerei der Universität.
Lindenberg, Siegwart (1992) "The method of decreasing abstraction." Pp. 3–20 in *Rational Choice Theory. Advocacy and Critique*, edited by James S. Coleman and Thomas J. Fararo. Newbury Park, CA: Sage.
Lindenberg, Siegwart (2001) "Social rationality versus rational egoism." Pp. 635–668 in *Handbook of Sociological Theory*, edited by Jonathan H. Turner. New York: Kluwer.
Lorenz, Edward H. (1988) "Neither friends nor strangers: Informal networks of subcontracting in French industry." Pp. 94–107 in *Trust: Making and Breaking Cooperative Relations*, edited by Diego Gambetta. Oxford: Blackwell.
Macaulay, Stewart (1963) "Non-contractual relations in business." *American Sociological Review* 28: 55–66.
Macy, Michael W., and Andreas Flache (1995) "Beyond rationality in models of choice." *Annual Review of Sociology* 21: 73–91.
Macy, Michael W., and Andreas Flache (2002) "Learning dynamics in social dilemmas." *Proceedings of the National Academy of Sciences U.S.A.* 99: 7229–7236.
Macy, Michael W., and Andreas Flache (2009) "Social dynamics from the bottom up. Agent-based models of social interaction." Pp. 245–268 in *The Oxford Handbook of Analytical Sociology*, edited by Peter Hedström and Peter Bearman. Oxford: Oxford University Press.
Mas-Colell, Andreu, Michael D. Whinston, and Jerry R. Green (1995) *Microeconomic Theory*. New York: Oxford University Press.
Merton, Robert K. (1968) *Social Theory and Social Structure*. Enlarged ed., New York: Free Press.
Merton, Robert K. (1973) *The Sociology of Science*, Chicago, IL: University of Chicago Press.
Milgrom, Paul, Douglass C. North, and Barry R. Weingast (1990) "The role of institutions in the revival of trade: The law merchants." *Economics and Politics* 2: 1–23.
Van Miltenburg, Nynke (2015) *Cooperation under Peer Sanctioning Institutions. Collective Decisions, Noise, and Endogenous Implementation*. PhD thesis, Utrecht University.
Van Miltenburg, Nynke, Vincent Buskens, and Werner Raub (2012) "Trust in triads: Experience effects." *Social Networks* 34: 425–428.
Nash, John F. (1951) "Non-cooperative games." *Annals of Mathematics* 54: 286–295.

North, Douglass C. (1990) *Institutions, Institutional Change and Economic Performance*. Cambridge: Cambridge University Press.
Olson, Mancur (1965) *The Logic of Collective Action*. 2nd ed., Cambridge, MA: Harvard University Press 1971.
O'Neill, John, ed. (1973) *Modes of Individualism and Collectivism*. London: Heinemann.
Opp, Karl-Dieter (1979) *Individualistische Sozialwissenschaft [Individualistic Social Science]*. Stuttgart: Enke.
Ortmann, Andreas, John Fitzgerald, and Carl Boeing (2000) "Trust, reciprocity, and social history: A re-examination." *Experimental Economics* 3: 81–100.
Ostrom, Elinor (2003) "Towards a behavioral theory linking trust, reciprocity, and reputation." Pp. 19–79 in *Trust and Reciprocity. Interdisciplinary Lessons from Experimental Research*, edited by Elinor Ostrom and James Walker. New York: Russell Sage.
Parsons, Talcott (1937) *The Structure of Social Action*. New York: Free Press.
Popper, Karl R. ([1934] 1973) *Logik der Forschung*. 5th ed. Tübingen: Mohr.
Popper, Karl R. (1945) *The Open Society and Its Enemies*. 2 vol., London: Routledge.
Popper, Karl R. (1957) *The Poverty of Historicism*. London: Routledge.
Popper, Karl R. (1963) *Conjectures and Refutations*. London: Routledge.
Popper, Karl R. (1972) *Objective Knowledge*. New York: Oxford University Press.
Portes, Alejandro (1998) "Social capital: Its origins and applications in modern sociology." *Annual Review of Sociology* 24: 1–24.
Prendergast, Canice (1999) "The provision of incentives in firms." *Journal of Economic Literature* 37: 7–63.
Rabin, Matthew (1993) "Incorporating fairness into game theory and economics." *American Economic Review* 83: 1281–1302.
Rapoport, Anatol (1974) "Prisoner's Dilemma – Recollections and observations." Pp. 18–34 in *Game Theory as a Theory of Conflict Resolution*, edited by Anatol Rapoport. Dordrecht: Reidel.
Rapoport, Anatol, Andreas Diekmann, and Axel Franzen (1995) "Experiments with social traps IV." *Rationality and Society* 7: 431–441.
Rasmusen, Eric (2007) *Games and Information: An Introduction to Game Theory*. 4th ed., Oxford: Blackwell.
Raub, Werner (1984) *Rationale Akteure, institutionelle Regelungen und Interdependenzen [Rational Actors, Institutional Rules, and Interdependencies]*. Frankfurt a.M: Lang.
Raub, Werner (1988) "Problematic social situations and the 'large-number dilemma'." *Journal of Mathematical Sociology* 13: 311–357.
Raub, Werner (1992) "Eine Notiz über die Stabilisierung von Vertrauen durch eine Mischung von wiederholten Interaktionen und glaubwürdigen Festlegungen [A note on stabilizing trust through a combination of repeated interactions and credible commitments]." *Analyse & Kritik* 14: 187–194.
Raub, Werner (1997) *Samenwerking in duurzame relaties en sociale cohesie [Cooperation in Durable Relations and Social Cohesion]*. Inaugural lecture, Utrecht University.
Raub, Werner (2004) "Hostage posting as a mechanism of trust: Binding, compensation, and signaling." *Rationality and Society* 16: 319–366.
Raub, Werner (2013) *Kwaliteit en kwantiteit in de wetenschap: Het transitieprobleem in analytisch perspectief [Quality and Quantity in Science: The Transitionproblem in Analytical Perspective]*. Mimeo, Utrecht University.
Raub, Werner (2014) "A theory of co-teaching." Pp. 145–159 in *Theorie en praktijk van leren en de leraar. Liber Amicorum Theo Wubbels*, edited by Jan van Tartwijk, Mieke Brekelmans, Perry den Brok, and Tim Mainhard. Amsterdam: SWP.
Raub, Werner (2016) *Een internationale sfeer in Langeveld, Groenman en Basket [An International Climate in Langeveld, Groenman, and The Basket]*. Mimeo, Utrecht University.
Raub, Werner, and Vincent Buskens (2006) "Spieltheoretische Modellierungen und empirische Anwendungen in der Soziologie [Game-theoretic modeling and empirical applications in sociology]." Pp. 560-598 in *Kölner Zeitschrift für Soziologie und Sozialpsychologie / Sonderheft 44: Methoden der Sozialforschung*, edited by Andreas Diekmann.
Raub, Werner, and Vincent Buskens (2008) "Theory and empirical research in analytical sociology: The case of cooperation in problematic social situations." *Analyse & Kritik* 30: 689–722.
Raub, Werner, Vincent Buskens, and Marcel A.L.M. van Assen (2011) "Micro-macro links and microfoundations in sociology." *Journal of Mathematical Sociology* 35: 1–25.

Raub, Werner, Vincent Buskens, and Rense Corten (2014) "Social dilemmas and cooperation." Pp. 597–626 in *Handbuch Modellbildung und Simulation*, edited by Norman Braun and Nicole J. Saam. Wiesbaden: Springer VS.

Raub, Werner, Vincent Buskens, and Vincenz 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, Werner, and Gideon Keren (1993) "Hostages as a commitment device: A game-theoretic model and an empirical test of some scenarios." *Journal of Economic Behavior and Organization* 21: 43–67.

Raub, Werner, and Chris Snijders (1997) "Gains, losses, and cooperation in social dilemmas and collective action: The effects of risk preferences." *Journal of Mathematical Sociology* 22: 263–302.

Raub, Werner, and Thomas Voss (1981) *Individuelles Handeln und gesellschaftliche Folgen [Individual Actions and Societal Level Implications]*. Darmstadt: Luchterhand.

Raub, Werner, and Thomas Voss (1986) "Die Sozialstruktur der Kooperation rationaler Egoisten [The social structure of cooperation of rational egoists]." *Zeitschrift für Soziologie* 15: 309–323.

Raub, Werner, and Thomas Voss (1990) "Individual interests and moral institutions. An endogenous approach to the modification of preferences." Pp. 81–117 in *Social Institutions: Their Emergence, Maintenance and Effects*, edited by Michael Hechter, Karl-Dieter Opp, and Reinhard Wippler. New York: Aldine.

Raub, Werner, and Thomas Voss (2017) "Micro-macro models in sociology: Antecedents of Coleman's diagram." Forthcoming in *Social Dilemmas, Institutions, and the Evolution of Cooperation. Festschrift for Andreas Diekmann*, edited by Ben Jann and Wojtek Przepiorka. Berlin: De Gruyter.

Raub, Werner, and Jeroen Weesie (1990) "Reputation and efficiency in social interactions: An example of network effects." *American Journal of Sociology* 96: 626–654.

Raub, Werner, and Jeroen Weesie (1992) *The Management of Matches*. Mimeo, Utrecht University.

Van de Rijt, Arnout, Soong Moon Kang, Michael Restivo, and Akshay Patil (2014) "Field experiments of success-breeds-success dynamics." *Proceedings of the National Academy of Sciences U.S.A.* 111: 6934–6939.

Van de Rijt, Arnout, and Michael W. Macy (2009) "The problem of social order: Egoism or autonomy?" *Advances in Group Processes* 26: 25–51.

Riker, William H., and Peter C. Ordeshook (1973) *An Introduction to Positive Political Theory*. Englewood Cliffs, NJ: Prentice-Hall.

Rooks, Gerrit, Werner Raub, Robert Selten, and Frits Tazelaar (2000) "Cooperation between buyer and supplier: Effects of social embeddedness on negotiation effort." *Acta Sociologica* 43: 123–137.

Rooks, Gerrit, Werner Raub, and Frits Tazelaar (2006) "Ex post problems in buyer-supplier transactions: Effects of transaction characteristics, social embeddedness, and contractual governance." *Journal of Management and Governance* 10: 239–276.

De Ruijter, Esther (2005) *Household Outsourcing*. PhD thesis, Utrecht University.

Sally, David (1995) "Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1992." *Rationality and Society* 7: 58–92.

Schelling, Thomas C. (1960) *The Strategy of Conflict*. London: Oxford University Press.

Schelling, Thomas C. (1978) *Micromotives and Macrobehavior*. New York: Norton.

Schneider, Louis, ed. (1967) *The Scottish Moralists on Human Nature and Society*. Chicago, IL: Phoenix.

Schotter, Andrew (1981) *The Economic Theory of Social Institutions*. Cambridge: Cambridge University Press.

Schüßler, Rudolf A. (1989) "Exit threats and cooperation under anonymity." *Journal of Conflict Resolution* 33: 728–749.

Selten, Reinhard. 1965. "Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit [A game-theoretic model of oligopoly with delayed demand]." *Zeitschrift für die gesamte Staatswissenschaft* 121: 301–324, 667–689.

Shaked, Avner (2006) *On the Explanatory Value of Inequity Aversion Theory*. Working paper.

Singer, Tania, and Anita Tusche (2014) "Understanding others: Brain mechanisms of theory of mind and empathy." Pp. 513–532 in *Neuroeconomics*, edited by Paul W. Glimcher and Ernst Fehr. 2nd ed., Amsterdam: Elsevier.

Snijders, Chris (1996) *Trust and Commitments*. PhD thesis, Utrecht University.

Stigler, George J. (1964) "A theory of oligopoly." *Journal of Political Economy* 72: 44–61.

Stinchcombe, Arthur L. (1975) "Merton's theory of social structure." Pp. 11–33 in *The Idea of Social Structure. Papers in Honor of Robert K. Merton*, edited by Lewis A. Coser. New York: Harcourt.
Taylor, Michael ([1976] 1987) *The Possibility of Cooperation*. Cambridge: Cambridge University Press (Revised edition of *Anarchy and Cooperation*. London: Wiley 1976).
Udehn, Lars (2001) *Methodological Individualism*. London: Routledge.
Vanberg, Viktor (1975) *Die zwei Soziologien. Individualismus und Kollektivismus in der Sozialtheorie [The Two Sociologies. Individualism and Collectivism in Social Theory]*. Tübingen: Mohr.
Vanberg, Viktor (1978) "Kollektive Güter und kollektives Handeln [Collective goods and collective action]." *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 30: 652–679.
Vanberg, Viktor J. and Roger D. Congleton (1992) "Rationality, morality, and exit." *American Political Science Review* 86: 418–431.
Vieth, Manuela (2009) *Commitments and Reciprocity. Experimental Studies on Obligation, Indignation, and Self-Consistency*. PhD thesis, Utrecht University.
Voss, Thomas (1982) "Rational actors and social institutions: The case of the organic emergence of norms." Pp. 76–100 in *Theoretical Models and Empirical Analyses. Contributions to the Explanation of Individual Actions and Collective Phenomena*, edited by Werner Raub. Utrecht: ESP.
Voss, Thomas (1985) *Rationale Akteure und soziale Institutionen [Rational Actors and Social Institutions]*. München: Oldenbourg.
Voss, Thomas (2001) "Game-theoretical perspectives on the emergence of social norms." Pp. 105–136 in *Social Norms*, edited by Michael Hechter and Karl-Dieter Opp. New York: Russell Sage.
Voss, Thomas (2015) *Der Beitrag der (experimentellen) Spieltheorie zur Sozialtheorie [The Contribution of (Experimental) Game Theory to Social Theory]*. Mimeo, Leipzig University.
Voss, Thomas (2016) *Institutional Design and Human Motivation*. Mimeo, Leipzig University.
Weber, Max ([1921] 1976) *Wirtschaft und Gesellschaft*. 5th ed., Tübingen: Mohr.
Weber, Max (1947) *The Theory of Social and Economic Organization*. New York: Free Press.
Weesie, Jeroen, and Werner Raub (1996) "Private ordering: A comparative institutional analysis of hostage games." *Journal of Mathematical Sociology* 21: 201–240.
Williamson, Oliver E. (1985) *The Economic Institutions of Capitalism*. New York: Free Press.
Williamson, Oliver E. (1996) *The Mechanisms of Governance*. New York: Oxford University Press.
Wippler, Reinhard (1978) "The structural-individualistic approach in Dutch sociology: Toward an explanatory social science." *Netherlands Journal of Sociology* 14(2): 135–155.
Wippler, Reinhard (1996) "Theoretische sociologie: Balans van een werkprogramma [Theoretical sociology: Taking stock]." Pp. 95–111 in Reinhard Wippler, *Sociologie tussen Empirie en Theorie. Een Keuze uit het Werk*. Amsterdam: Thela Thesis.
Wippler, Reinhard, and Siegwart Lindenberg (1987) "Collective phenomena and rational choice." Pp. 135–152 in *The Micro-Macro Link*, edited by Jeffrey C. Alexander, Bernhard Giesen, Richard Münch, and Neil J. Smelser. Berkeley, CA: University of California Press.
Wittek, Rafael, Tom A.B. Snijders, and Victor Nee, eds. (2013) *Handbook of Rational Choice Social Research*. Stanford, CA: Stanford University Press.
Wrong, Dennis H. (1961) "The oversocialized conception of man in modern sociology." *American Sociological Review* 26: 183–193.
Yamagishi, Toshio, and Midori Yamagishi (1994) "Trust and commitment in the United States and Japan." *Motivation and Emotion* 18: 129–166.
Ziegler, Rolf (1972) *Theorie und Modell [Theory and Model]*. München: Oldenbourg.
Van der Zwaan, Bert (2017) *Higher Education in 2040*. Amsterdam: Amsterdam University Press.

Contents in detail

PhD students

Abraham, Martin (1995) *Betriebliche Sozialleistungen und die Regulierung individueller Arbeitsverhältnisse. Endogene Kooperation durch private Institutionen*. University of Erlangen-Nuremberg; supervision with Günter Büschges; present position: full professor, University of Erlangen-Nuremberg.

Snijders, Chris (1996) *Trust and Commitments*. Utrecht University; supervision with Jeroen Weesie and Tom A.B. Snijders; present position: full professor, Eindhoven Technical University.

Flache, Andreas (1996) *The Double Edge of Networks. An Analysis of the Effect of Informal Networks on Cooperation in Social Dilemmas*. Groningen University; supervision with Wim B.G. Liebrand and Frans N. Stokman; present position: full professor, Groningen University.

Blumberg, Boris F. (1997) *Das Management von Technologiekooperationen. Partnersuche und Verhandlungen mit dem Partner aus empirisch-theoretischer Perspektive*. Utrecht University; supervision with Frits Tazelaar, Jeroen Weesie, and Reinhard Wippler; present position: assistant professor, Maastricht University.

Prosch, Bernhard (1998) *Die Absicherung von Lieferbeziehungen. Partnersuche, vertragliche Festlegungen und soziale Einbettung beim Einkauf von EDV-Produkten*. Leipzig University; supervision with Thomas Voss and Günter Büschges; final position: assistant professor and Privatdozent, University of Erlangen-Nuremberg (deceased).

Buskens, Vincent (1999) *Social Networks and Trust*. Utrecht University; supervision with Jeroen Weesie; present position: full professor, Utrecht University.

Van Assen, Marcel (2001) *Essays on Actor Models in Exchange Networks and Social Dilemmas*. Groningen University; supervision with Frans N. Stokman and Tom A.B. Snijders; present position: full professor, Utrecht University, and assistant professor, Tilburg University.

Jansen, Miranda (2002) *Waardenoriëntaties en Partnerrelaties. Een Panelstudie naar Wederzijdse Invloeden*. Utrecht University; supervision with Matthijs Kalmijn and Aat C. Liefbroer; present position: managing director of the Division Social Sciences and Methods and Statistics, Utrecht University.

Rooks, Gerrit (2002) *Contract en Conflict. Strategisch Management van Inkooptransacties*. Utrecht University; supervision with Frits Tazelaar and Chris Snijders; present position: assistant professor, Eindhoven Technical University.

Gautschi, Thomas (2002) *Trust and Exchange. Effects of Temporal Embeddedness and Network Embeddedness on Providing and Dividing a Surplus*. Utrecht University; supervision with Chris Snijders and Jeroen Weesie; present position: full professor, University of Mannheim.

De Ruijter, Esther (2005) *Household Outsourcing*. Utrecht University; supervision with Tanja van der Lippe and Jeroen Weesie; present position: senior researcher, AO Consult, Tilburg.

Koster, Ferry (2005) *For the Time Being. Accounting for Inconclusive Findings Concerning the Effects of Temporary Employment Relationships on Solidary Behavior of Employees*. Groningen University; supervision with Karin Sanders and Frans N. Stokman; present position: full professor, TIAS, Tilburg University, and associate professor, Erasmus University Rotterdam.

Barrera, Davide (2005) *Trust in Embedded Settings*. Utrecht University; supervision with Vincent Buskens; present position: assistant professor, University of Turine.

De Valk, Helga (2006) *Pathways into Adulthood. A Comparative Study on Family Life Transitions among Migrant and Dutch Youth*. Utrecht University; supervision with Aat C. Liefbroer and Pearl A. Dykstra; present position: full professor, Groningen University and Free University Brussels, senior researcher, Netherlands Interdisciplinary Demographic Institute (NIDI).

Vogt, Sonja (2007) *Heterogeneity in Social Dilemmas: The Case of Social Support*. Utrecht University; supervision with Vincent Buskens and Jeroen Weesie; present position: senior research officer, Nuffield College, University of Oxford.

Knecht, Andrea (2008) *Friendship Selection and Friends' Influence: Dynamics of Networks and Actor Attributes in Early Adolescence*. Utrecht University; supervision with Tom A.B. Snijders, Chris Baerveldt, and Christian Steglich; present position: policy advisor, City of Nuremberg.

Van Houten, Gijs (2008) *Beleidsuitvoering in Gelaagde Stelsels. De Doorwerking van Aanbevelingen van de Stichting van de Arbeid in het CAO-Overleg*. Utrecht University; supervision with René Torenvlied; present position: senior researcher, Eurofound, Dublin.

Vieth, Manuela (2009) *Commitments and Reciprocity. Experimental Studies on Obligation, Indignation, and Self-Consistency*. Utrecht University; supervision with Jeroen Weesie and Vincent Buskens; present position: postdoc, University of Basel.

Corten, Rense (2009) *Co-evolution of Social Networks and Behavior in Social Dilemmas*. Utrecht University; supervision with Vincent Buskens and Stephanie Rosenkranz; present position: associate professor, Utrecht University.

Roeters, Anne (2010) *Family Life under Pressure? Parents' Paid Work and the Quantity and Quality of Parent-Child and Family Time*. Utrecht University; supervision with Tanja van der Lippe and Esther Kluwer; present position: senior researcher, The Netherlands Institute for Social Research (Sociaal Cultureel Planbureau, SCP).

Mäs, Michael (2010) *The Diversity Puzzle. Explaining Clustering and Polarization of Opinions*. Groningen University; supervision with Andreas Flache, Rafael Wittek, and Karoly Takacs; present position: assistant professor, Groningen University.

Zhelyazkova, Asya (2012) *Compliance under Controversy. Analysis of the Transposition of European Directives and their Provisions*. Utrecht University; supervision with René Torenvlied and Robert Thomson; present position: postdoc, ETH Zürich.

Schalk, Jelmer (2012) *The Performance of Public Corporate Actors. Essays on Effects of Institutional and Network Embeddedness in Supranational, National, and Local Collaboration Contexts*. Utrecht University; supervision with René Torenvlied and Agnes Akkerman; present position: assistant professor, Leiden University.

Bojanowski, Michal (2012) *Essays on Social Network Formation in Heterogeneous Populations. Models, Methods, and Empirical Analyses*. Utrecht University; supervision with Vincent Buskens and Stephanie Rosenkranz; present position: senior researcher, Interdisciplinary Center for Mathematical and Computational Modelling (ICM), University of Warsaw.

Minescu, Anca (2012) *Relative Group Position and Intergroup Attitudes in Russia*. Utrecht University; supervision with Louk Hagendoorn and Edwin Poppe; present position: assistant professor, University of Limerick, Ireland.

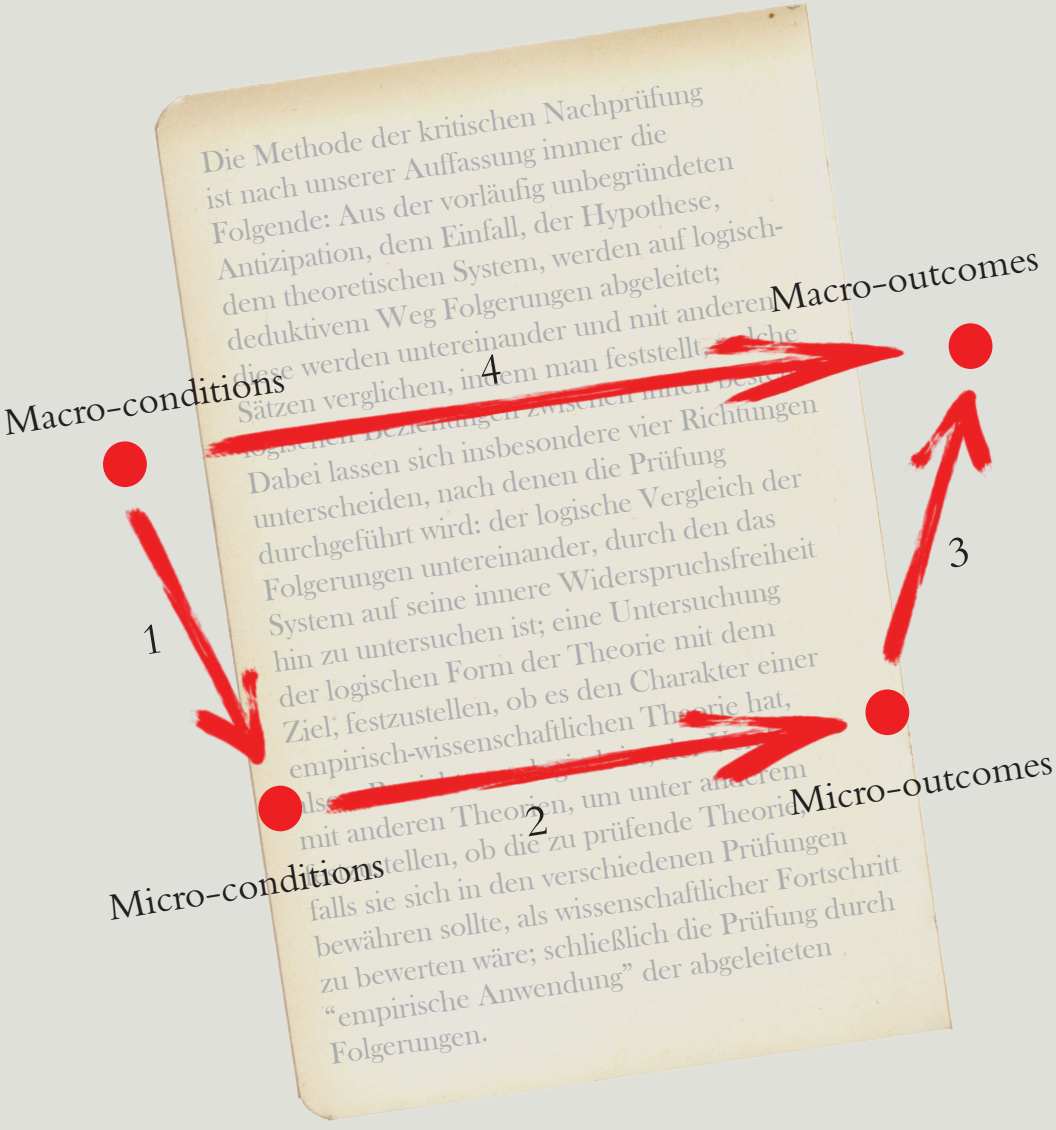
Van Schellen, Marieke (2012) *Marriage and Crime over the Life Course. The Criminal Careers of Convicts and their Spouses*. Utrecht University; supervision with Paul Nieuwbeerta and Anne-Rigt Poortman; present position: assistant professor, Utrecht University.

Aksoy, Ozan (2013) *Essays on Social Preferences and Beliefs in Non-Embedded Social Dilemmas*. Utrecht University; supervision with Jeroen Weesie; present position: senior lecturer, University College London.

Morbitzer, Dominik (2013) *Limited Farsightedness in Network Formation*. Utrecht University; supervision with Vincent Buskens and Stephanie Rosenkranz; present position: researcher at Marketing Science, Ipsos, Munich.

Van Miltenburg, Nynke (2015) *Cooperation under Peer Sanctioning Institutions. Collective Decisions, Noise, and Endogenous Implementation*. Utrecht University; supervision with Vincent Buskens and Davide Barrera; present position: senior analyst, Willis Towers Watson, Zurich.

Frey, Vincenz (2016) *Network Formation and Trust*. Utrecht University; supervision with Vincent Buskens and Rense Corten; present position: postdoc, Bocconi University, Milan.



Werner Raub is Professor of Theoretical Sociology at Utrecht University and the Interuniversity Center for Social Science Theory and Methodology (ICS). He has been Dean of Social and Behavioral Sciences at Utrecht University from 2012 to 2017.

Rational Models is on three topics: sociology as a problem- and theory-guided discipline; rational choice models of trust and cooperation in social dilemmas, including empirical research in this field; implications for running academic institutions.