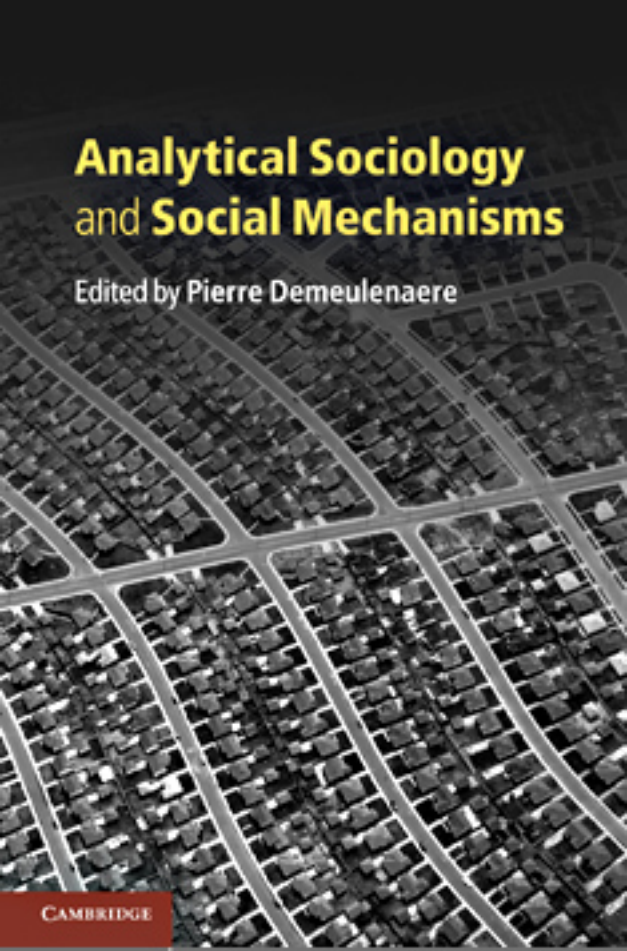
Analytical Sociology and Social Mechanisms

Edited by Pierre Demeulenaere



CAMBRIDGE

CAMBRIDGE www.cambridge.org/9780521190473

This page intentionally left blank

Analytical Sociology and Social Mechanisms

Mechanisms are very much a part of social life. For example, we can see that inequality has tended to increase over time, and that cities can become segregated. But how do such mechanisms work? Analytical sociology is an influential approach to sociology which holds that explanations of social phenomena should focus on the social mechanisms that bring them about. This book evaluates the major features of this approach, focusing on the significance of the notion of mechanism. Leading scholars seek to answer a number of questions in order to explore all the relevant dimensions of mechanism-based explanations in social sciences. How do social mechanisms link together individual actions and social environments? What is the role of multi-agent modelling in the conceptualization of mechanisms? Does the notion of mechanism solve the problem of relevance in social sciences explanations?

PIERRE DEMEULENAERE is Professor of Sociological Theory and Philosophy of the Social Sciences at the University of Paris-Sorbonne.

Analytical Sociology and Social Mechanisms

Edited by

Pierre Demeulenaere

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town,
Singapore, São Paulo, Delhi, Tokyo, Mexico City

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press,
New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521154352

First published 2011
Printed in the United Kingdom at the University Press, Cambridge

A catalogue record for this publication is available from the British Library

Library of Congress Cataloguing in Publication data
Analytical sociology and social mechanisms / [edited by]
Pierre Demeulenaere.
p. cm.
ISBN 978-0-521-15435-2 (pbk.)
1. Sociology. 2. Sociology–Methodology. 3. Social systems.
I. Demeulenaere, Pierre. II. Title.
HM585.A526 2011
301.01–dc22
2010052189

ISBN 978-0-521-19047-3 Hardback
ISBN 978-0-521-15435-2 Paperback

This volume was published with the support of the Université Paris-
Sorbonne.

Contents

Figures

Tables

Contributors

PETER ABELL, Copenhagen Business School and London School of Economics.

YVONNE ÅBERG, Stockholm University.

RAYMOND BOUDON, Institut de France, Académie des sciences morales et politiques.

DAMON CENTOLA, Massachusetts Institute of Technology.

PIERRE DEMEULENAERE, University of Paris-Sorbonne.

JON ELSTER, Collège de France.

THOMAS J. FARARO, University of Pittsburgh.

ANDREAS FLACHE, University of Groningen.

PETER HEDSTRÖM, Nuffield College, University of Oxford.

MICHAEL W. MACY, Cornell University.

GIANLUCA MANZO, Centre National de la Recherche Scientifique, Paris and University of Paris Sorbonne.

ROBERT J. SAMPSON, Harvard University.

KEITH SAWYER, Washington University in St Louis.

MICHAEL SCHMID, Bundeswehr Munich University.

DAN SPERBER, Centre National de la Recherche Scientifique, Paris.

ARNOUT VAN DE RIJT, State University of New York at Stony Brook.

ROBB WILLER, University of California at Berkeley.

PETRI YLIKOSKI, University of Tampere and University of Helsinki.

Introduction

Pierre Demeulenaere

Why should we introduce the notion of "analytical sociology" into the field of sociology, and why should it be linked to the concept of "mechanism"?

I do not believe there to be any great need to introduce new paradigms into a discipline already encumbered with so many antagonizing trends, schools and paradigms. Analytical sociology should not therefore be seen as a manifesto for one particular way of doing sociology as compared with others, but as an effort to clarify ("analytically") theoretical and epistemological principles which underlie any satisfactory way of doing sociology (and, in fact, any social science). The social sciences already command a considerable stock of substantive descriptions and explanations; and some of the alternatives to these are either redundant, or resistant to proof, even false or imprecise, quite regardless of their status with respect to one or other established paradigm. Analytical sociology should seek to define a set of sound epistemological and methodological principles underlying all previously established and reliable sociological findings. The aim of analytical sociology is to clarify the basic epistemological, theoretical and methodological principles fundamental to the development of sound description and explanation.

The recurrent use of the term "analytical" in sociology derives mainly from the accepted notion of "analyticity," designating a division into basic elements, the difficulty being in the determination of these clear-cut basic elements, since such division is not universally accepted – recently the notion of "holism" has been associated with a refusal to accept such a separation (see, for example, Demeulenaere 2000; Descombes 1996). For instance, the constitutive elements of a belief cannot be precisely separated in the same way that two actors can be isolated from one another. Even when we separate one actor from another, the fact that his beliefs depend to a great extent on previously acquired knowledge means that he cannot be completely separated from the environment in which such knowledge has been acquired. This is

why any attempt to separate these elements must coincide with epistemological reflection on the relevance of such an "analytical" enterprise. The most important aspect of the analytical approach should be to clarify the strategy by which we endeavour to separate and conceptualize different elements entering into descriptions and explanations of the social world, so that we might understand their mutual relationships, and in particular the causal links existing among them.

The use of the notion of "analyticity" relates first of all to emphasis upon the idea that any description or explanation necessarily involves separate "elements" to be considered in respect of their specificity, status and role. This separation leads on to an elucidation of the manner in which they are reciprocally articulated, and in particular are said to "cause" one another. This is why the "mechanism" issue is necessary to any explanation. Whenever we start explaining "why" something happens, beyond mere description, we are necessarily led to introduce some type of causal linkage of elements that in turn raises the question of mechanism. Analytical sociology is impelled in this way toward the study of mechanisms and their functioning. Emphasis on the notion of mechanism corresponds to an evaluation of the proper role of causal linkages in the social sciences. But as we shall see later in this introduction, there are many mutually antagonistic views of the notion of causality, its role in sociological explanation, and its relation to the notion of mechanism. One aim of this volume is to clarify the relationships arising between these various uses and conceptions.

Where does this notion of analytical sociology come from? It is common to find the use of the adjective "analytic" in the social sciences, emphasized to a greater or lesser degree. Among the major theorists, Talcott Parsons is notable for development of the notion of an "analytical" approach in sociology. His aim was to discover and isolate abstract features of the social world (see Fararo 1989, and also Chapter 5 in this volume, for an overview of Parsons' contribution to the ideal of analytical theory in sociology; Fararo himself uses the notion of "analytical action").

But more recently, and in a different manner, the term has been reintroduced into social science literature and also given broader scope. This was in the book edited by Peter Hedström and Richard Swedberg and published in 1998 with the title *Social Mechanisms*, and whose subtitle was: *An Analytical Approach to Social Theory*. Peter Hedström subsequently published an important short book called *Dissecting the Social. On the Principles of Analytical Sociology* (2005). This book is a systematic exposition of what can be called analytical sociology. A handbook was then published outlining a general program of research (Hedström and

Bearman 2009). A synthesis had previously appeared in Italian, outlining the principal features of such a trend (Barbera 2004). Moreover, several papers have now addressed the key issues of this new movement. It has prompted vigorous debate at an international level (Manzo 2010).

This volume presents a collection of chapters dealing with central issues raised by some of the most important authors in this movement. This does not mean that all the contributors consider themselves to be part of a single movement; nor that this movement is a perfectly unified school united by common and consistent beliefs. The idea is to discuss and clarify the main issues involved in such an enterprise from an epistemological, methodological and theoretical viewpoint. The book is not a manifesto either pro or contra analytical sociology and the use of mechanisms: it is an attempt to reflect upon the key issues involved and in particular the use of the notion of causality in sociological explanation.

Analytical sociology and methodological individualism

Since social theory is still very often associated with scholars who have defined principles and theories, we can start by evoking the theorists who can be included in a list of "analytical thinkers":

1. First, some classical authors, such as Tocqueville (Hedström and Edling 2009) or Merton (Hedström and Udéhn 2009) are considered in retrospect by current analytical sociologists to exemplify analytical sociology in principle. More generally, any classical author who has advanced a convincing explanation of social phenomena with a clear understanding of the social mechanisms at work can in retrospect be considered an analytical sociologist. It is important to note that Boudon (1998), for example, has consistently sought to reconsider the work of mutually opposing authors so that he might demonstrate a deeper underlying unity in their arguments.
2. Second, Hedström considers some of the most important writers of modern social science – Schelling, Coleman, Boudon and Elster – to be the contemporary founding fathers of analytical sociology. Since these four authors are commonly presented (or have presented themselves) as "methodological individualists," the link between methodological individualism and analytical sociology has to be addressed. Both Boudon and Elster have accepted the label of "analytical sociologists" by publishing papers in books seeking to define analytical sociology, the present collection included. By contrast,

there are those such as Arthur Stinchcombe who, while contributing to the theory of social mechanisms, do not appear to be direct participants in the movement.

3. Finally, many contributors to books such as the present one can be seen to either support the movement, be interested in its core issues, or associate themselves with its main debates.

Whatever the case may be, there is a clear connection between the tradition of methodological individualism (MI) and the rise of analytical sociology (AS). The four contemporary authors Hedström considers to exemplify this approach are usually classified as methodological individualists. Critics of AS include those who are equally critical of MI. Therefore, the question of the link that exists between the two movements is to be analyzed.

The two core ideas behind MI, first expressed by John Stuart Mill and Carl Menger, and subsequently by Weber, can be expressed very simply:

1. Social life exists only by virtue of actors who live it.
2. Consequently a social fact of any kind must be explained by direct reference to the actions of its constituents.

These two simple propositions remain central to the analytical approach; we therefore have to address the problem of the relationship between MI and AS. This section is directed to a brief exposition of the problem.

To sum up the main features of MI I will begin with a foundational quotation from one early proponent of the approach. MI can be shown to be at variance with the principles claimed for current analytical sociology or, conversely, can be shown to be substantively similar to and continuous with these principles.

> According to this principle, the ultimate constituents of the social world are individual people who act more or less appropriately in the light of their dispositions and understanding of their situation. Every complex social situation, institution or event is the result of a particular configuration of individuals, their dispositions, situations, beliefs, and physical resources and environment. There may be unfinished or half-way explanations of large-scale social phenomena (say, inflation) in terms of other large-scale phenomena (say, full employment); but we shall not have arrived at rock-bottom explanations of large-scale phenomena until we have deduced an account of them from statements about the dispositions, beliefs, resources, and interrelations of individuals. (Watkins 1957, 1959: 505)

Taking this early statement as a simple example of a definition of what MI sought to be, it can be said that it is generally oriented to three major misunderstandings; that is, not objections regarding its relevance, but

criticisms arising from misconceptions regarding what defenders of MI generally stated in their writings. I will later argue that an emphasis on AS is a way of avoiding such misconceptions, and generating a broader consensus concerning what "good social science" should be.

The first of these misconceptions is the claim that MI is "atomistic." This assertion goes back to an old dispute between nineteenth-century economists (in particular Menger (1996 [1883]) who first developed this notion of atomism) and advocates of sociology or the social sciences who insisted that individuals were never isolated, but were instead dependent on their social environment, which environment was often called a "structure" (although this is a notion open to many interpretations). Granovetter (1985) famously restated this criticism of "atomism" in order to introduce the idea of an "embedded" actor. However, it should be clear in the quotation from Watkins that there is no intrinsic link between MI and atomism. As Homans puts it:

> The position taken makes no assumption that men are isolated individuals. It is wholly compatible with the doctrine that human behavior is now and always has been social as long as it has been human. (Homans 1967: 59)

Hence two positions are compatible with the aim of methodological individualists, and both can be derived from Watkins' quotation:

1. Actors depend, in their behavior, on interrelation with others, the resources they possess, and the institutions in which such behavior evolves.
2. Beliefs and motives are founded upon knowledge and on norms which are both social in this sense, and which are not of their own making. Therefore actors evolve in a cultural and social environment, defining their objectives and representations in terms of this environment. Accordingly, a reference to "rock-bottom" explanations does not imply "atomistic" or non-social actors, but instead evokes "dispositions, beliefs, resources and interrelations of individuals" as opposed to macro-social laws. Watkins does go on to write that MI should be contrasted with "holism," which is itself contrasted with a "rock-bottom" explanation:

   > On this latter view, social systems constitute "wholes" at least in the sense that some of their large-scale behaviour is governed by macro-laws, which are essentially sociological in the sense that they are *sui generis* and not to be explained as mere regularities or tendencies resulting from the behaviour of interacting individuals. (Watkins 1957, 1959: 505)

It should be clear that the refusal to adopt a macro-law perspective does not in any sense imply the assumption of non-social actors. The

rejection of macro-laws is not at all equivalent to a refusal of the "socialness" of actors. Macro-laws and desocialized individual atoms are not alternatives. There is therefore absolutely no reason that MI should be seen as a device separating actors from their environment, reducing explanation to "individual" features or actors and consequently annihilating any reference to their environment. There is a tendency to confuse individual actors with dissocialized actors.

On the contrary, this environment has to be taken into account if we are to understand "individual" actions. We can cite Homans again here:

> Sociologists do not often realize that they pursue two related, but often distinguishable subjects for empirical research. Most sociologists pursue one far more often than they do the other; a few pursue both. The first, which I shall call *individualistic* sociology, is concerned with the way in which individuals in interaction with one another create structures, and the second which I shall call *structural* sociology, is concerned with the effects these structures, one created and maintained, have on the behaviour of individuals or categories of individuals. In the empirical propositions of the former, the behaviour of individuals is treated as the set of independent variables and the characteristics of the structures as the set of dependent ones. In the latter the process is reversed: the structures are treated as the set of independent variables and the behaviour of individuals as the set of dependant ones. (Homans 1984: 341)

The combination of these two approaches can be called "structural individualism" (Udéhn 2001; Wippler 1978). Any serious attempt to reflect on a social situation should deploy both in turn. Their combination is in some respect illustrated by Coleman's famous "boat" (1986, 1990). It remains a central aspect of analytical sociology. I will however come back later to this difficult and central issue of the opposition between macro and micro levels, an issue which has been revived within MI and analytical sociology.

A second frequent misrepresentation of MI is to assume it to be utilitarian. Clearly, some authors in the MI tradition have, more or less explicitly, held utilitarian positions – for instance, Homans (1967), Coleman (1990) or Hechter and Opp (2001) amongst others. But some major theorists reject such an association (Boudon 2001, or Elster 2009 for instance). The appeal of utilitarianism derives from the difficulty of understanding any action not oriented to gaining some kind of advantage (from the point of view of the actor). In this sense, even suicide is a remedy for a life gone wrong. But this notion of an "advantage" is imprecise, open to many varied and contrasting constructions. Whenever we try to define in a more precise way the exact content of utilitarian motives we encounter a dilemma: either they are specified

narrowly, and so then quite plainly do not correspond to the broad range of human motives; or they are so loosely specified that they cover all individual preferences (rooted in social contexts), the notion of utility then losing its own specificity and becoming redundant, since any kind of preference becomes part of a utility function (Hollis 1994; Sen 1977). We should not therefore reduce MI to a narrow version of utilitarianism, whether from a descriptive viewpoint (since many authors do not support such a position) or from a normative viewpoint, since this dilemma stands in the way of any such reduction.

Another misunderstanding follows on from this: the conflation of MI with a narrow form of "rational choice" theory. This involves four different problems:

1. First, the very definition of the notion of "rational" behavior is at issue. What exactly should this notion of rationality imply: perfect information? Transitivity of preference orders? Intentionality? The choice of solution to a problem? All of these are widely debated, and there is no clear consensus on the meaning of rationality. Nevertheless, since MI authors constantly emphasize individual actions, we must take into account the intentional dimension of action, and also therefore its link with the notion of rationality.
2. Second, the normative dimension of rationality can be perceived as a problem in need of elimination from scientific discourse. Homans (1987) for instance argued that it was unnecessary to introduce such normative concepts into sociology or psychology. However, it is also possible to argue that, since human behavior is intrinsically normative, normative concepts should necessarily be central to any scientific analysis of such behavior. Weber was well-known for adopting this position. Normativity and the way in which actors deal with it must itself be explained, since the world, and individual action in particular, has normative features. As Joseph Raz puts it, "the core idea is that rationality is the ability to realise the normative significance of the normative features of the world, and the ability to respond accordingly" (Raz 2000: 35).
3. Third, there is the problem of the link between intentional action (purposive action) and emotion; this has been most notably discussed by Elster (1999). Whenever we stress the possibility of irrational behavior we need to consider – beyond simple description or normative assessment of irrational behavior – the conditions under which an actor is likely to act as either a rational or an irrational actor. Not everybody indulges in wishful thinking; hence the difficult question from the standpoint of rationality is

understanding why some act in a rational manner whereas others do not. A related question is whether wishful thinking has to be opposed to rational decision-making, since it might appear to be supported by some form of evidence.

4. Finally, the problem of habit on the one hand, and of creativity on the other, can be added to the above (Gross 2009); for it should seem obvious that people very often act on the basis of unreflecting habit. Should this be treated as a challenge to a theory of intentional action, or be on the contrary integrated into it? In my view, emphasizing habit or creativity should not lead us to a radical break with the idea of intentional actions, since habits are often presented as pragmatic solutions to problems. But this is clearly an issue for a debate. Similarly, emphasizing actions instead of actors (Abbott 2007) does not significantly alter the problem of interpreting the way in which action occurs, and should be so interpreted.

The third major misunderstanding about MI stems from the very notion of individualism. To what exactly does it correspond? As I said above, the core simple idea of MI is that there is no social life without so-called "individuals" being its motivating agents. The word "individuals" is clearly misleading here. It should not mean that these agents are separated from their environment, or that they necessarily act on the basis of "selfish" motives. Is there nevertheless an additional, specifically "individual" dimension of the actors that should be taken into account whenever an explanation is provided? It seems to me that two different things should be simultaneously stated. They appear to conflict, but they can be reconciled by stratifying the level of analysis.

First of all, referring to individual actors does not necessarily imply a reference to strictly naturalistic pre-social (and in this sense "individualistic") motives. For instance, in Schelling's famous example (1978), the actors' preferences for a relatively mixed neighborhood are given as social preferences regarding individuals. They could be significantly different. Culture should not therefore be seen to be absent from micro-level explanations. It is not because we refer to "individual"-level explanations that culture, in all its richness and complexity, is set to one side. Bearman *et al.*'s (2004) study of the sexual and romantic networks of adolescents in a midwestern American high school refers to a set of norms necessary to an understanding of the adolescents' choices, although these norms are not actually articulated by these adolescents. The principal norm is one prohibiting "from a boy's point

of view, that he formed a partnership with his prior girlfriend's current boyfriend's prior girlfriend" (Hedström and Bearman 2009). This norm permits explanation of the network structure under consideration, but is not explained in itself. It clearly has a cultural dimension (we are not, for instance, in a situation where sexual relationships between boys and girls are forbidden outside marriage, but at the same time relations are obviously not completely free). Therefore, any reference to a "micro" level, or to a so-called "individualistic" level, does not necessarily entail that "culture" is set aside. By contrast, when Coleman illustrates the micro level of analysis by using Weber's famous example, citing the manner in which actors endorsing Protestant values are led to specific economic attitudes, he clearly introduces a cultural dimension at the micro level. There is therefore no necessary opposition between individualism and the "socialness" of the actor (Little 2009: 163), since the notion of individuals can encompass a variety of cultural features.

That said, reference to a variety of cultural settings should not necessarily be the last word in social scientific explanation. Culture and its norms should also be treated as social facts to be explained, and not treated as something beyond the scope of further investigation (Mantzavinos 2005). The norms that constitute culture, varying significantly from one context to another, can themselves be seen to be enigmas in need of elucidation. In so doing, we have to move toward more universal motives and forms of actions which, combined with particular settings, can explain the prevalence of certain types of norms in certain situations. There is always a pressure in the social sciences to find, behind the existence of cultural and social diversity, some common features of human behavior which allow us to explain this diversity. It is not an easy task, and it may often fail, but the logic of such an effort depends upon the identification of relatively stable motives and attitudes so that we might understand the real variety of diverse motives and attitudes. In so doing, we inevitably tend to presuppose a kind of "human nature" representing general features of the species. These can therefore be called "individual" (although this is a rather misleading term), insofar as every human being globally reflects these general features of the human species, even if all singular individuals do not actually resemble each other and do display different features. These common features of human action are not necessarily non-social, since, for instance, the ability to find solutions to problems involves cooperation and discussion. Another explanatory move can lead to infra-individual causal determination, beyond conscious intentional actions (Chapter 3, in this volume).

From methodological individualism to analytical sociology

Does analytical sociology differ significantly from the initial project of MI? I do not really think so. But by introducing the notion of analytical sociology we are able to make a fresh start and avoid the various misunderstandings now commonly attached to MI. AS retains the core MI idea (that all social events depend on so-called "individual" actions which are responsible for the realization of social phenomena), but puts to one side all the misconceptions attached to "individuals," while also focusing attention on the complexity of those theoretical, epistemological and methodological issues involved in sociological explanations, and in particular that of the causality linking social events.

Since all social life involves "individual" actors, and any explanation of the social world requires reference to their actions, analytical sociology turns on a remodeled theory of action, having two main dimensions.

First, a redefinition of the general features of action constituting social life, irreducible to some narrow form of rational action. Hedström has proposed a so-called Desire–Belief–Opportunity theory of action, which can be derived from Hume's theory of motivation as developed by the analytical tradition of philosophy. It can be found for instance in the work of Elizabeth Anscombe, described by Mark Platts in the following terms:

> Miss Anscombe, in her work on intention, has drawn a broad distinction between two kinds of mental states, factual belief being the prime exemplar of one kind and desire a prime exemplar of the other … The distinction is in terms of the direction of fit of mental states with the world. Beliefs aim at the true and their being true is their fitting the world; falsity is a decisive failing in a belief, and false beliefs should be discarded; beliefs should be changed to fit with the world, not *vice versa*. Desires aim at realisation, and their realisation is the world fitting with them; the fact that the indicative content of a desire is not realised in the world is not yet a failing in the desire, and not yet any reason to discard the desire; the world, crudely, should be changed to fit with our desires, not vice versa. (1979: 256–7. Quoted in Smith 1994: 111–12)

The problem of normativity arises when a distinction between belief and desire is developed, since a belief has no normative strength of its own. The introduction of normative beliefs (or "besires") thus creates conceptual problems (Smith 1994). Sociology cannot afford to ignore discussion of these in the literature of analytical philosophy, for seeking explanation of social norms (and beliefs regarding those norms) implies reflection on their source, or their normativity (Dancy 2000).

In any case, analytical sociologists are interested in the theory of action involved in social science explanation. The relation between desires, beliefs and opportunities is to be linked to the general issue of meaning (intentionality and the interpretation of this action in terms of rationality or irrationality).

Meaning also involves the cultural settings in which individual actions find their relevance. Let me comment here on an extremely simple sentence, taken from Chapter 10 in this volume: "young people are less likely to use umbrellas than older people are" (see p. 203). This can be analyzed as a simple correlation between two states. But were it to be "understood," it could be deduced either from a cultural feature (adolescent culture does not favor the use of umbrellas, unless they are fashionable) or from a psychological dimension (young people are less bothered by rain and cold). A causal link between the two dimensions (youth psychology leading to youth culture) can be conceived. Homans' emphasis on behavioral psychology is today certainly outdated. This does not mean that any reference to psychology (through individual representations and emotions) should be discarded: this is why renewed attention to psychological and cognitive constraints can be seen as a major task of analytical sociology. But at the same time, this does not exclude attention to cultural and "social" meanings: these are certainly not irreconcilable and should not be superficially opposed, but rather consistently, if tentatively, linked. At any rate, an emphasis on "desire" does not exclude its link to culture: desires can be very "natural," such as the desire for sexual intercourse, or very "cultural," such as the desire to wear a white wedding dress.

Second, the promotion of structural individualism (the notion of which is prior to the current development of AS and is, in my view, inherent to sociological MI from the very beginning, as opposed to some versions of economic atomism) is based on the fact that actors act in a social environment, and that the environment "plays a role" in their decisions. But methodological individualists have always defended the idea that individuals are, let us say, "embedded" in social situations that can be called "social structures," and are in no respect isolated atoms moving in a social vacuum. But "playing a role" is certainly a rather elusive notion, since we seek to maintain simultaneously two things that seem to be opposites. On the one hand, institutions and rules depend on actors to be active; whereas that which is "natural" is defined as such through its possession of its own motive forces. Institutions and rules have no direct "energy" of their own. But on the other hand, when we say that they "play a role" we refer to the fact that their existence (however conceived) clearly has effects upon individual

action. How we might explain this circumstance is a central question for analytical sociology. It is clear for instance that the notion of "path dependency" establishes an understandable connection from the fact of an institution being established at one point in time to the subsequent difficulty for any movement toward radical institutional change (North 1990). Hence we commonly encounter the fact that, one set of actors having at a certain point adopted a norm, or built up an institution, it then becomes more difficult for actors subsequently coming onto the scene to adopt another norm, or replace the institution. But this is never impossible.

Mechanisms and causality

The notion of mechanism is therefore a key issue for analytical sociology. Why is this so? Responding to this question involves two rather different things.

The first involves a general regard for causality and its role in social scientific explanation. Analytical sociology incorporates an affirmation that "social facts" are generated, triggered, produced, brought about or "caused" by individual actions which themselves are in some sense "caused," or at least partly determined by the constraints presented by the social environments and situations in which such actions take place. To explain a social event therefore means to describe the various causal chains linking all the elements involved (once those elements have been appropriately described and separated) in constituting a social fact. This also means identifying the relevant elements between which causal relationships exist, and determining their nature. From this perspective, a mechanism is the set of elements and their causal links that regularly lead from an initial social state to a subsequent one.

But beyond this general regard for causality, we can observe an additional and more specific introduction of the notion of "mechanism" distinct from, and narrower than, a mere causal relationship. The conception of "causal regularities" can fail to provide sufficient explanatory traction, whereas the notion of mechanism permits a more certain grasp of the principal relations involved. The introduction of the notion of mechanism was initially intended as an alternative to the inductive regularity thesis (Little 1991). It was moreover understood to be an attack on Hume's conception of causality as well as on Hempel's covering-law theory of explanation. Both theories are naturally very different; and the covering-law approach is not properly speaking a theory of causality, since it criticizes this notion. Nevertheless, a theory

of mechanisms has been regarded as an alternative to both of these accounts of what causality is and what causality represents in the social sciences.

I will try to provide here a brief account of ongoing debates regarding the significance of the idea of mechanism for social science explanation in which analytical sociologists have been involved. My intention is certainly not to present a solution, but to outline briefly the types of discussion associated with the notion of mechanism.

There are two distinct dimensions in the implication of mechanisms that should not be confused:

1. The first relocates the effective locus of causal links in social science explanation from the level of macro-social variables to that of micro-social action.
2. The second discusses the true nature of causality and challenges other theories of causality (namely those of Hume or Hempel) by introducing a specific meaning associated with the idea of mechanism.

The two questions are distinct: the first locates real causal relationships in social scientific explanation, and the second assesses, on philosophical grounds, what a causal relationship actually is.

I will not here recapitulate the entire history of mechanism-based explanations, in particular in economics and in mathematical sociology, which are certainly prior to the recent development of analytical sociology (see Cherkaoui 2005 and Berger 2010 for an overview). I will mainly emphasize the causality issue involved.

First of all, the link between explanation and causality arises from an attempt to answer the question of "why" something occurs. In responding to the question we are inevitably led to introduce causal patterns whenever we use the term "because." As Elster puts it, "all explanation is causal explanation. We explain an event by citing its cause. Causes precede their effects in time" (Elster 2007: 271). In other words, it is said that any explanation should refer to causal links between elements. This view can be challenged (Kitcher 1989; Psillos 2002) Nonetheless, it is essential that the social sciences reflect upon the use of such notions in "causing," "producing," "developing," "generating," "triggering," or "bringing about" social outcomes. The constant use of such terms in sociological discourse without seeking to illuminate their conditions of applicability and validity is very unsatisfactory. One of the great achievements of sociologists associated in some way with the analytical sociology movement has been to abandon this unreflecting attitude toward causality, so that its usage might be better controlled.

Before developing this question, it should be said that the argument concerning an intrinsic link between strategies of explanation and causal relations is incomplete, since there is at least one other another widespread and significant use of "explanation" in the social sciences which should not be overlooked: where the explanation of an action or an object is equivalent to giving its "meaning" (for the different uses of the notion of explanation, see Salmon 1998: 5). For instance, to explain why someone does something we can refer his gesture to its meaning in a definite cultural setting, or the connection it has with other meaningful intentions. The resulting answer does not imply that the culture or the set of motives has "caused" the gesture. It only specifies the connections that such a gesture has with other meaningful elements constitutive of the culture. This dimension of the concept of "explanation" should not be excluded from AS because, as argued above, AS certainly does not exclude the cultural dimension of an action, nor should it do so.

However, the specific emphasis on social mechanisms seeks to provide explanations that illuminate the causal links existing between elements. As shown above, one central aspect of methodological individualism was its refusal to recognize the validity of causal chains posited between variables at a macro level, for the very simple reason that causal connections do not genuinely occur at this level, the real motivation for any change being located in human action, which is necessarily a micro-level event. Causal relations should therefore be rooted in the effective actions clearly responsible for any one event, even though these are constrained by their environment and the manner in which such action is institutionally organized and "structured." For instance, if a higher average speed is correlated with a higher average number of accidents on a highway, it can be said "metaphorically" that higher average speed "causes" a higher number of accidents. But it is obvious that the relevant causality involves cars and their drivers, and it is out of the behavior of drivers and the cars under their control that such abstract elements as average speed and average number of accidents are constructed, and related one to another. Moreover, each accident necessarily implies locally a complex set of additional causal relations where speed is only one element amongst others. This is based on a common-sense notion that an "average speed" has no causal power as such, whereas one particular driver traveling at a very high speed is able to trigger an accident. We refer to such a common-sense notion of causality when we know, for instance, that "the number 2 cannot engage in a process that turns it into the number 3: only a brain can now think of 2, now of 3. Likewise, there is no mechanism whereby the word 'dogs'

transforms itself into the word 'gods': only a brain can now think of dogs, now of gods" (Bunge 2004: 374). Thus opposition to the interpretation of variable correlations in terms of straightforward causality is based on two distinct considerations:

1. The first involves the possibility of false positives and false negatives: the fact that, on the one hand, statistical correlations do not necessarily involve causal links, and on the other, that true causal relationships are sometimes not revealed by statistical correlations (because of the intervention of other factors) (Little 1991: 24).
2. The second concerns the necessity of a focus upon those actions responsible for any change. This does not involve a sophisticated theory of what a causal link is. It only leads negatively to a refusal to interpret simple correlation between variables as a causal link, since macro variables have no causal power by themselves. To repeat Watkins' point, "there may be unfinished or half-way explanations of large-scale social phenomena (say, inflation) in terms of other large-scale phenomena (say, full employment); but we shall not have arrived at rock-bottom explanations of large-scale phenomena until we have deduced an account of them from statements about the dispositions, beliefs, resources, and interrelations of individuals."

Following the logic of such an idea, any interpretation of variables in terms of causal links would seem defective, since it omits the mechanisms triggering the link. In Fararo's words:

> In the context of present-day empirical research, a great deal of stress is placed on the specification of variables and explaining variation in one variable by appeal to variations in a battery of other variables. This conception of explanation is based on a legitimate aspect of the dynamic or process-oriented frame of reference of general theoretical sociology. Namely … as the generative mechanism produces changes of state, it does so under parametric conditions. Variations in parametric conditions, either dynamically or comparatively, thereby produce variations in states *via* the generative mechanisms. What is omitted in the usual accounts of "causal models" based on "explaining variation" is the explicit formal representation of process. (Fararo 1989: 42)

Therefore the focus has to be on the causal "process" occurring at the action level. The idea that there are laws directly implemented at a macro level can be easily rebutted, since the effectiveness of the outcome necessarily leads to the "active" level, the level of action. One general implication of the notion of "mechanism" is to move analysis away from an "inactive" level to an "active" level, where effective actions occur. A strong correlation between variables should not therefore be interpreted in causal terms unless a mechanism linking the two dimensions

is identified, mechanism involving effective actions. Should the correlation of variables thus be considered causal once the mechanism linking them has been revealed? This is in part a matter of definition, since causality derives from other effective causal connections. There would be two different levels of causality, direct and indirect (although the direct level necessarily involves quite different other causal dimensions which permit their realization).

Two contrasting views have nonetheless been held regarding the level of action itself: on the one hand, some sociologists have accepted the idea that there are laws operating at the level of action. A classical example can be found in Homans' theorization of social explanation. He explicitly accepts Hempel's theory of explanation based on laws. He sees here no difference between the natural and the social sciences. There are laws at the action level that allow us to provide explanations for specified behavior. Homans gives examples of explanations of this kind which perfectly illustrate Hempel's principles:

> Take, for instance, the proposition that strongly patrilineal societies are apt to have institutionalized a close, warm, end free relationship between a man and his mother's brothers. Once societies had adapted patrilinity for whatever reasons in the past, the forces thus set in motion must have been strong enough to make the societies converge towards the ego-mother's brother relationship, just as societies that industrialized set in motion forces tending towards a nuclear family system. Crudely and elliptically put the argument is as follows: that in patrilineal societies fathers hold jural authority over their sons; that authority, because it restricts freedom and may be exercised through punishment, tends to inhibit a close, warm, and free relationship between the person in authority and the person submitted to it; that a man may nevertheless have strong needs that can be satisfied through such a relationship, especially with an older person of the same sex; and, finally, that the nearest older man that does not have authority over him, is a man's mother's brother.
>
> Note that the argument is inherently psychological. More important for my present purposes, it is broadly genetic. (Homans 1967: 101)

This position is also held by Opp (2005). By contrast, other sociologists or philosophers reject the idea of an explanation based on covering laws at the micro level because it is said not to explain the effective "mechanism" at work in the phenomena to be explained. The first step was to introduce mechanism-based explanation in order to move beyond the (false, or inappropriate) assumption that the simple occurrence of correlation between variables could be interpreted in terms of causality. It is the mechanism which allows us to demonstrate the process at work when the outcome is produced through effective actions. But a second and distinct step was introduced regarding the very meaning of causality at this "lower level." It led to the idea that Hempel's

covering-law principle is equally inadequate at this level, and that it should be replaced by the intervention of a mechanism (for those who believe in such mechanisms). No longer is the focus on the fact that correlation between macro variables is an inadequate basis for a true appraisal of causality, but rather on the fact that even at the micro level the notion of law should be replaced by the notion of "mechanism," since general laws are not easy to find.

There are therefore two completely distinct issues at stake. The first is relatively consensual (Abbott, for instance, supports this view): all interpretation in terms of causality of macro-variable explanations based on variable correlations is considered unsatisfactory unless the action level mechanisms triggering the social phenomena can be identified. The second questions the very existence of causal links at the level of action itself, with two alternative views: one that would replace causal links by "mechanisms," posing then the problem of the relation between causality and mechanisms at the action level (Little 2009); and another which would abandon the very idea of causality and/or of mechanisms. In other words, when we turn away from thinking in terms of correlations between variables and toward the idea of "mechanisms" at the individual action level, we then need to determine the elements responsible for social change at the level of actions. Are there laws at the level of action? If we replace them by the notion of mechanism, how does it articulate with causality? And if we were to abandon the notion of causality, how would an explanation be possible?

This raises the problem of the nature of causality itself and its relation to the notion of a "causal mechanism." It is inevitable that reflection upon causal links in social scientific explanation should lead into an epistemological investigation of how we might characterize the existence of an effective causal relationship. Determining what it is that allows us to speak of causality is a general philosophical question. But more modestly, it is important to differentiate two quite distinct things.

First is the fact that mere correlation is not a genuine explanation of a causal relationship, and that we need to identify intermediate causal steps at the level of action so that we might explicate the production of social outcomes. This position is based on the common-sense position that in social life only actions produce social change; it does not involve any metaphysical claim about the nature of a true causal relationship.

Second, this is not equivalent to replacing the notion of a causal link by the notion of a generative mechanism, since it can be said on the one hand that a simple correlation between variables is not a genuine causal link, and on the other that mechanisms are necessarily based upon causal links. In other words, emphasizing the importance of generative

mechanisms as a methodological explanatory strategy representing an alternative to statistical correlations is not necessarily equivalent to the replacement of a so-called "succession" theory of causation by a "generative" theory, the way Rom Harré describes the contrast between those two approaches:

> The two great metaphysical theories of causality take their difference from the way they treat the relation between the cause and its effect. In the *generative* theory the cause is supposed to have the power to generate the effect and is connected to it. In the *succession* theory a cause is just what usually comes before an event or state, and which comes to be called its cause because we acquire a psychological propensity to expect that kind of effect after the cause. The difference between the theories can be expressed in other ways. For the generative theory the relation between the events or states or happenings that are related is internal to them, the cause and the effect could not happen without the cause. It would not be just what it is were it differently caused. And part of what it is to be the happening which generates a certain effect. On the other hand, the succession theory treats the causal relation as external to the cause and effect so related. The cause can be described perfectly and completely without reference to what effect it has and the effect of a cause is an independently specifiable event or happening or state which would be just what it is had it been spontaneously generated. (Harré 1972: 116)

Elster similarly criticizes covering-law based explanations, employing slightly different arguments:

> The two problems we have just discussed add up to a weakness in the best-known theory of scientific explanation, that proposed by Carl Hempel. He argues that explanation amounts to logical deduction of the event to be explained, with general laws and statements of initial conditions as the premises. One objection is that the general laws might reflect correlation, not causation. Another is that the laws, even if genuinely causal, might be preempted by other mechanisms. This is why I have placed the emphasis here on mechanisms, not on laws. (Elster 1989: 6)

So, both authors insist on a "generative" dimension of causality as opposed to a pure "succession" theory. They do not distinguish the macro-variables level and the action level. For the sake of clarity, we might note that this line of argument involves two different things.

The first corresponds to the contrast between an alleged causal relationship linking macro variables and the true generative process involving effective actions. It could however be said that in this respect there is no truly effective causal linkage between variables at the macro level, since the linkage necessarily depends on the effective actions of individuals. The causal link is not demonstrated. This does not necessarily challenge Hume's theory of causality based as it is on constant and regular succession, which only stresses the fact that in order to be able

to acknowledge the very existence of a causal link we have no other basis than an admission that we have a *constant* repetition of a succession of events. The problem with correlations between macro variables is that, on the one hand, we have no proof that this is a constant relation (so we are not sure that we are in a true "succession" situation); and on the other, that we know that the macro variable depends on actors for its realization: therefore to identify the effective causal link we necessarily have to investigate the actions which are part of the process. In other words, the interpretation of variable correlations in terms of causality can be criticized in Hume's terms, and equally in Hempel's terms (although such criticisms will be very different), without reference to a "generative" theory.

Conversely, a reference to "mechanism" does imply stable connections, which necessarily correspond to the stable succession of events as required by Hume or Hempel.

This is why Elster actually plays down his opposition to Hempel's form of explanation:

> This is not a deep philosophical disagreement. A causal mechanism has a finite number of links. Each link will have to be described by a general law, and in that sense by a "black box" about whose internal gears and wheels we remain ignorant. Yet for practical purposes – the purposes of working social scientists – the place of emphasis is important. By concentrating on mechanisms, one captures the dynamic aspect of scientific explanations: the urge to produce explanations of ever finer grain. (Elster 1989: 6–7)

I believe that Harré likewise does in fact refer to regular causal links when he describes a mechanism:

> To explain a phenomenon, to explain some pattern of happenings, we must be able to describe the causal mechanism which is responsible for it. To explain the catalytic action of platinum we must not only know in which cases platinum does catalyse a chemical reaction, but what the mechanism of catalysis is. To explain the fact of catalysis we need to know or to be able to imagine a plausible mechanism for the action of catalysis. Ideally a theory should describe what really is responsible for whatever process we are trying to understand. But this ideal can rarely be fulfilled. In practice, it becomes this: ideally a theory should contain the description of a plausible iconic model, modelled on some thing material, or process which is already well understood, as a model of the unknown mechanism, capable of standing in for it in all situations
>
> …
>
> If knowledge is pursued according to this method it will tend to be stratified. Perhaps this can be seen best if we look at the way causes are elaborated. There are two conditions which have to be fulfilled for there to be truly said to be a causal relation among happenings or phenomena. The fist condition, ensuring that there is prima facie evidence, is that there should seem to be some pattern

or structure in what we observe to be happening. This might be that simple kind of pattern which we call regularity or repetition, when we find one sort of happening followed regularly by happenings of a certain, definite other kind, when for instance those who are deprived of fresh fruit and vegetables develop scurvy, and those who have plenty of the above commodities do not. We have prima facie evidence that there is a causal relation between the deprivation and the disease. But to eliminate all possibility that something else, some third factor, might be responsible both for the shortage of vegetables and for the scurvy, we must find out what is the mechanism involved, and that involves us in a study of the chemistry of the food materials and of the physiology and chemistry of the body. That study supplies an idea of the mechanism which explains the patterns of happenings involving presence and absence of fresh vegetables and the onset and cure of scurvy. Satisfying this second condition, that is, describing the causal mechanisms completes one causal study. Our knowledge falls out into two strata as it were: in one stratum the facts to be explained are set out and their pattern described; in the underlying stratum we may imagine or describe the causal mechanism.

Now, that mechanism is described in terms of chemical reactions and physiological mechanism. These exhibit their own characteristic patterns and regularities, and these call again for causal explanation. But now a new kind of facts must be adduced. Chemical reactions are explained by the theory of atoms and molecules and chemical valency. By means of this model we can describe a causal mechanism for chemical reactions, and similar considerations apply to the explanation of the physiological and biochemical facts. We have reached another stratum. Then that stratum itself becomes the occasion for prima facie hypotheses that there are too causal relations, that there is some mechanism which explains the combining powers of chemical atoms which would explain the diversity of chemical elements. (Harré 1972: 178–9)

In this remarkable passage, it is clear that any kind of mechanism does in effect ultimately rely upon causal regularities. But to be explained it is said that they should lead to a "lower level" mechanism. I will return to this important dimension in a moment.

Before pursuing this line of thought, I will just add that, seeking a complete description of the debate, Hedström and Udéhn, in their paper on Merton (2009) in the *Handbook*, introduce a fourth dimension of causality (in addition to the dimensions of covering law, statistical association and mechanisms). This is associated with Woodward's theory of counterfactuals, not mentioned in Hedström's 2005 book, and which is very possibly due to the influence of Petri Ylikoski (who presents Woodward's position in this volume, in Chapter 8). Briefly, the counterfactual dependencies theory outlined by Woodward (2003) emphasizes the consequences that interference with the factors said to be responsible for an event would have on the production of this event:

The fourth alternative, causal mechanisms, should not necessarily be seen as an alternative to the counterfactual approach, but rather as adding further requirement. As emphasized by Woodward (2002) and Morgan and Winship (2007) there is nothing in the counterfactual approach as such which guarantees sufficient causal depth, because perceptions of sufficient depth are discipline-specific while the counterfactual approach is not … There is nothing in the counterfactual approach which suggests that an explanation in terms of the macro–micro–macro path is preferable to the macro-macro path … As long as an intervention which changes a macro-property (the explanans) invariably leads to a change in another macro-property (the explanandum), a strict counterfactualist would be satisfied, although a relationship at that level is a black box which offers no information on why the properties are related to one another. In this respect, the causal-mechanisms alternative differs from the counterfactual alternative in that it insists that the link between the explanans and the explanandum should be expressed in terms of the entities and activities through which the relationship is believed to have been brought about, i.e. in this case in terms of individuals and their actions. (Hedström and Udéhn 2009: 42)

An important aspect of the theory of mechanisms is that it frequently refers to a hierarchy of levels, and introduces an explanatory "lower" level in respect of the higher level to be explained. This is clearly stated in Harré's above quotation. It is seen as a consensual aspect of the theory of mechanisms (Gross 2009). It is considered by Stinchcombe to be an essential aspect of social mechanisms:

The core of the definition of a mechanism is the definition of lower-level units of analysis which have causal unity. By this we mean that the units of analysis are places [or things] where a given kind of causal forces (I will usually call them *inputs* because we are interested in how the larger system functions and are using the smaller units of analysis to explain) are connected to their effects (I will usually call them *outputs*). "What are the causal unities that can form units of analysis?" Is an empirical [and therefore ultimately theoretical] question … I think the four main kinds of causal unities in the social sciences, making up units of analysis in which mechanisms operate are (1) individuals, (2) social actors that can be treated as individuals (such as firms), (3) situations, and (4) patterns of information [such as languages]. (Stinchcombe 1991, 1993: 28)

However, it might be noted that combining a "macro" level with a "superior" level is metaphorical; while commonly deployed, and especially in Coleman's theorization of the micro/macro link, the metaphor is nevertheless misleading. When we designate a "macro" level, we commonly refer to three rather different things which are all equally "social," since they cannot refer to one singular individual, but only to a group of individuals.

1. First, we refer to the statistical characteristics of a group, based on the distribution of individual features, such as average income, average age or the median.
2. Second, we refer to the relative positions different actors take up one to another (in a city, in a hierarchy, in a network, in a set of relations or resources and so on).
3. The third general meaning of a "macro" level corresponds to produced "social objects" with which individuals have to deal in their social life: these include social norms, shared beliefs and social institutions.

The combination of those three genuinely heterogeneous aspects makes up the "macro" environment, or "structure," in which one evolves: for instance cars on a highway first have an average speed (which is determined on the basis of the various individual speeds); second they are in different positions relative to each other, at close or remote distance (reinforced or not by appropriate distance norms); and third, the actors follow (or not) particular speed norms, and act in an institutional environment (the way the road is designed, the signs on the road that prescribe a maximum speed allowed, a minimum distance between cars to be respected, etc.).

Should all of these three dimensions be equally considered part of a "superior level," or of a "system level," in the way that Coleman describes things (1990)? This can be a matter of debate, since this spatial metaphor certainly does not give an accurate account of the various ways through which these three various "macro" dimensions articulate to the micro ones, that is, via the action of individuals: a mean is not relative to an individual in the same way as a set of relative positions, or as a rule which one can follow or not. In the first case, it is a collective measure based on a series of individual measures, but there is no social novelty that "emerges." In the second case, it is a pattern of relative positions, which is sometimes institutionalized, sometimes not; that is, it might be stable despite the substitution of individuals, or conversely it can disappear when individuals change or move. And in the third case corresponding to norms, institutions and shared beliefs, there are in effect new social objects that are relatively independent from individuals once they have been established. Individuals can activate them, or be constrained by them (through the action of other individuals).

Therefore should we similarly speak of an "emergent" system level, or alternatively of a "supervenient" level (Hedström and Bearman 2009), corresponding to the notion of a "superior" level associated with a "macro" level? It can be noticed that the frequent use of the notion

of emergence in the social sciences refers to two fundamentally heterogeneous meanings: one is an ontological idea of a "superior" level; the other is the temporal appearance of a new social object which was previously absent (the way for instance a norm is said to "emerge" from an iterated Prisoner's Dilemma situation, or an unforeseen consequence of an action appears in the course of its realization). On the one hand, using as a model what some describe as typical for the natural sciences, we refer to an emergent level when the components of a given reality are not considered sufficient to provide a full account of the reality at stake, which is in this sense said to be "superior" or "emergent." This was clearly Durkheim's line of argument regarding society (Sawyer 2005). Is this the case for the calculation of a mean, for a distributional set of positions, and for the appearance of a rule or an institution? It could be argued that in these cases nothing is really superior, and there is no emergent property at work in such situations.

On the other hand, when a norm is set up, or an institution developed, or new shared beliefs appear, these clearly are new objects introduced into social life: for instance a new bank, a new fashion, or a new currency. It can be said that they "emerge," although the ontological novelty is not at all the same as that previously conceptualized. We have new objects in the domain of social life, whose appearance is sometimes intended, sometimes not. Should we then say that these objects emerge at a "higher" "system" level? This would mean that the rule is "above" the micro level of actors who follow a rule: but there is no need here to adopt a strong notion of emergence based on a similarity with the natural sciences, where higher levels "emerge" from lower ones; the same can be said of supervenience. The rule does not supervene on individuals in the sense that consciousness supervenes on a neuronal substratum, since the rule can be disconnected from the individuals who adopt it or reject it. To sum up, when we refer to "macro" elements, we designate a heterogeneous set of elements, which can be said to "emerge" or to "supervene" in different ways, with different meanings attached to those notions.

Therefore explanation involves various types of relationship between the micro and the macro level: the influence of norms on actual behaviors, the necessity to discover new norms, the consequence of individual decisions on the way people decide to move or to remain in one area of a city, etc. There is no unity behind the macro elements involved and the relations they have with individuals. It is not therefore obvious to say that the constitutive element of a mechanism-based explanation is to locate a lower level for a higher level: what is important in social scientific mechanisms is reference to the active level, which is to the way in

which social change is produced. Therefore there is no necessary link, in social scientific explanations based on mechanisms, between a move from a higher to a lower level: these are metaphors designating various relations between properties concerning separate individuals and properties concerning groups of individuals, involving different types of causal links. Sometimes individuals directly influence a collective outcome, when a rule or an institution is set up. A macro novelty is produced through the action of individuals. Sometimes the relation is indirect: if one individual in a group moves from one neighborhood to another, there is no new macro property triggered in the previous sense, but an evolution from one distribution of proprieties to a new one.

An overview of the book

The contributors to the book include a range of major authors in contemporary sociology who treat different categories of objects, as discussed above. First of all, some contributors to this volume are prominent authors in the analytical tradition, such as Boudon and Elster, or Hedström who has been responsible for the theorization of this approach. Second, there are authors who, although not directly associated with the official history of the movement, have approached key issues of fundamental concern to it, such as Fararo's interest in generative mechanisms. The book does not aim to present a systematic and programmatic approach to the analytical movement, and the issue of mechanisms. Instead, it offers a series of reflections upon its main themes. It focuses in particular on the question of causality in relation to mechanisms in sociological explanation, and offers various points of view on this which derive from the epistemological difficulties involved.

The book is organized into three parts on the basis of the above discussion. Part I reassesses aspects of the theory of action (rationality, emotions, beliefs) that are important for sociological explanations. It also discusses the connections that actions have with their environment, at an "inferior" or at a "superior" level.

Part II introduces an extended reflection on the link between mechanisms and causality. It does so from the perspective of the formalization of generative models and also from an epistemological perspective.

Part III presents empirical examples interpreted in the light of mechanisms. It permits reflection upon the micro–macro link in empirical data analysis. It also includes a presentation of agent-based modeling, which is an important methodology for analytical sociology.

Part I introduces new aspects of action theory in the field of analytical sociology, and its relation to its context. In particular, it presents

Raymond Boudon's revised version of rationality to demonstrate the fact that analytical sociology has no necessary connection to a narrow version of rational choice. In Chapter 1, Boudon presents an alternative to rational choice theory which he calls a theory of ordinary rationality. Its aim is to understand and explain not only the means that are chosen to reach ends, but also the ends themselves and the values people endorse. The chapter proposes a formal definition of this notion of ordinary rationality. It shows that it can be applied to representations and values, presents a sample of its applications, and sketches a survey of its logical advantages over the theories of rationality currently in use, such as the so-called Rational Choice Theory (RCT) or the Theory of Bounded Rationality (TBR).

Jon Elster is another important figure in analytical sociology. In Chapter 2, Elster stresses the role of emotions and beliefs in a theory of action. He similarly departs from the standard rational choice model of action, where the role of beliefs is limited to informing the agent of the best means to realize his or her desires. The emotion-based model he proposes relies on the fact that beliefs can also generate emotions that have consequences for behavior. He discusses the specific link between beliefs and emotions through various examples. This chapter is an epistemological reflection upon the works Elster has already, in other volumes, devoted to this topic.

Dan Sperber's contribution is included in this volume to show the possible consequences of an emphasis on "causal" explanation in respect of the level at which the existence of causal connections should be acknowledged. In Chapter 3, Sperber pursues two related objectives. He departs from the narrow model of intentional action and introduces infra-individual elements causally determining behavior. Doing so, he provides arguments for a naturalistic ontology which is, in his view, the only basis for arriving at sound causal claims in social scientific explanation. He therefore proposes a naturalization of the domain of the social sciences, modeled not on psychology but on epidemiology.

To understand the role of social structure, in Chapter 4 Keith Sawyer considers (somehow conversely to the previous chapter) that mechanism-based explanations should not be solely related to individual actions, nor to an infra-individual level, but rather to an "emergent level": communicative interactions should be the fundamental units of the system. They are said to be "emergent" in the sense that they do not depend on the intentions of the individuals. Social mechanistic explanations should therefore introduce emergent forms and macro-social structures. The chapter usefully presents a theory of emergence related to mechanisms.

Part II is devoted to an analysis of the relationship existing between mechanisms and causality. Thomas Fararo is a key figure of mathematical sociology, and has promoted the notion of generative mechanism. He is therefore an important figure for analytical sociology. In Chapter 5, Fararo does not address causality as such, but the related concept of generativity and the modeling of generative processes. He starts with a description of the way the theories of Parsons and Homans lacked this generative dimension in their interpretation of social processes. He does so in reference to mathematical models, and proceeds to discuss generative models in the context of the development of cognitive science. The chapter develops thereafter a discussion of generativity in computational sociology, which forges a link to the two last chapters in this volume.

The issue of causality is, as we have seen, a difficult topic for the social sciences: it is often derived from the repetition of successive events; with an emphasis on the role one event has in producing another (the question being whether or not we need the repetition of successive events in order to acknowledge such a production). In Chapter 6 Peter Abell addresses the important question of causality when events are not repeated. Is it possible to speak of causality in the case of unique events when the idea of causality is not supported by the repetition of a succession of similar events? Can causal analysis be conceived when events and actions are not repeated? Abell suggests that causal inference in situations of this sort must rely upon a singular concept of causality. Strings (chronologies) of actions can be depicted as narratives, the causal links of which are singular. He introduces the analysis of such singular causal links in terms of Bayesian narratives.

Michael Schmid endeavors in Chapter 7 to describe the multi-level mechanistic character of explanation in the social sciences. We have seen in the Introduction that there are problems related to the definition of the notion of levels and of their relations. Schmid addresses this problem of level and their articulation, and argues that social explanation should do without reference to macro laws, without however being reduced entirely to an individual level. He then describes four steps in the explanation: the first step occurs at the individual level and involves causally determined individual actions; the second step consists in determining how the different actors link their action with each other; the third step then identifies the "collective" consequences of common co-adjustment attempts. Then the fourth step identifies the recursive effects of such co-adjustments on individual actions.

In Chapter 8 Petri Ylikoski addresses directly the notion of mechanism and its importance for social scientific explanation. This notion is

commonly introduced, in part as a solution to the relevance problem raised by covering laws, emphasized notably by Salmon. But, Ylikoski argues, a reference to mechanisms does not provide a satisfactory solution to the problem of explanatory relevance. Explanatory relevance should instead be achieved by introducing the so-called manipulationist account of causation which gives a sense of causal explanation. It relies on counterfactuals: an explanatory claim is correct if ideal intervention on the *explanans* variable brings the appropriate change in the *explanandum* variable. The chapter explores the consequences of this idea for explanation in the social sciences.

Finally in Part II, in Chapter 9, Pierre Demeulenaere explores the links between causal regularities in ordinary actions and the explanation of such actions: there is no explanation in social sciences that does not rely on practical regularities. This therefore dispenses with a sharp contrast between mechanism-based explanations and covering-law explanations, since they both formally rely on regularities and causal links. Nonetheless, should those regularities be unilaterally called laws? The answer has to be nuanced, because of the importance of institutions and rules for action integrated into causal explanations. The logic of explanation formally obeys Hempel's scheme so that it might show causal connections, but the rules to which it refers often do not have the stability of natural laws. We do therefore have the possibility of causal explanations without at the same time having strictly natural laws.

Finally, Part III is devoted to empirical investigations of social phenomena that can be interpreted through mechanisms. In Chapter 10, Yvonne Åberg and Peter Hedström first define what a social interaction effect is, and then use unique population-level panel data to show the importance of social interactions for explaining youth unemployment levels between different neighborhoods. The mechanisms here are the social interaction effects on the levels of unemployment.

Robert Sampson analyses neighbourhood effects. The results summarized in Chapter 11 support the notion that neighborhood selection is part of a process of stratification that encompasses individual decisions made within an ordered, yet constantly changing, residential landscape. Selection and sorting are conceptualized as part of a dynamic social process of neighborhood stratification that reproduces racially shaped economic hierarchies and that leads to an apparently durable equilibrium. Sorting is at once a causal mechanism and a social process.

Finally, two chapters present the developing methodology of computational models which exemplifies important developments in analytical sociology. In Chapter 12, M. Macy *et al.* describe the logic and

significance of such models. Agent-based computational models are *agent-based* because they take as a theoretical starting point a model of the autonomous yet interdependent individual units (the "agents") that constitute a dynamic system. The models are *computational*, because the individual agents and their behavioral rules are formally represented and encoded in a computer program such that the dynamics of the model can be deduced using step-by-step computation from given starting conditions.

In Chapter 13 Gianluca Manzo presents the theoretical structure and computational results of an agent-based model built in NetLogo language, which contains two groups of mechanisms:

1. generative mechanisms of proportions of dissatisfied actors;
2. generative mechanisms of the intensity of actors' feelings of dissatisfaction.

The aim is at the same time substantive and methodological. From a methodological point of view, this preliminary analysis aims to suggest that agent-based modeling constitutes a particularly well-adapted tool for analytical sociology, enabling sociologists in this field to study as completely and rigorously as possible how posited mechanisms work and produce aggregate and individual effects. At the same time, it attempts to establish a link between the notion of causality and the methodology of computational models.

REFERENCES

Abbott, Andrew. 2007. "Mechanisms and relations," *Sociologica* 2.

Barbera, Filippo. 2004. *Meccanismi Sociali; Elementi di sociologia analitica*. Bologna: Il Mulino.

Bearman, Peter S., J. Moody and K. Stovel. 2004. "Chains of affection: the structure of adolescent romantic and sexual networks," *American Journal of Sociology* 110: 44–91.

Berger, Nicolas. 2010. "Sociologie analytique, mécanismes et causalité: Histoire d'une relation complexe," *L'Année Sociologique* 60(2): 421–43.

Boudon, Raymond. 1998. *Etudes sur les sociologues classiques*. Paris: P.U.F.

2001. *The Origin of Values*. New Brunswick and London: Transaction.

Bunge, Mario A. 1997. "Mechanism and explanation," *Philosophy of the Social Sciences* 27: 410–65.

2004. "Clarifying some misunderstandings about social systems and their mechanisms," *Philosophy of the Social Sciences* 34(3): 371–81.

Cherkaoui, Mohamed. 2005. *Invisible Codes*. Oxford: Bardwell Press.

Coleman, James S. 1986. "Social theory, social research, and a theory of action," *American Journal of Sociology* 91: 1309–35.

1990. *Foundations of Social Theory*. Cambridge and London: The Belknap Press of Harvard University Press.

Dancy, Jonathan (ed.) 2000. *Normativity*. Oxford: Blackwell.
Demeulenaere, Pierre. 2000. "Individualism and holism: new controversies in philosophy of social sciences," *Mind and Society* 2(1): 3–16.
Descombes, Vincent. 1996. *Les institutions du sens*. Paris: Editions de Minuit.
Elster, Jon. 1989. *Nuts and Bolts for the Social Sciences*. Cambridge University Press.
1999. *Alchemies of the Mind: Rationality and the Emotions*. Cambridge University Press.
2007. *Explaining Social Behaviour. More Nuts and Bolts for the Social Sciences*. Cambridge University Press.
2009. *Le désintéressement. Traité critique de l'homme économique*. Paris: Seuil.
Fararo, Thomas J. 1989. *The Meaning of General Theoretical Sociology. Tradition and Formalization*. Cambridge University Press.
Granovetter, Mark. 1985. "Economic action and social structure: the problem of embeddedness," *American Journal of Sociology* 91(3): 481–510.
Gross, Neil. 2009. "A pragmatist theory of social mechanisms," *American Sociological Review* 74: 358–79.
Harré, Rom. 1972. *The Philosophies of Science. An Introductory Survey*. Oxford University Press.
Hechter, Michael and Karl-Dieter Opp. 2001. "What have we learned about the emergence of norms?" in Michael Hechter and Karl-Dieter Opp (eds.), *Social Norms*. New York: Russell Sage Foundation, 394–416.
Hedström, Peter. 2005. *Dissecting the Social. On the Principles of Analytical Sociology*. Cambridge University Press.
Hedström, Peter and Peter Bearman. 2009. *The Oxford Handbook of Analytical Sociology*. Oxford University Press.
Hedström, Peter and Christopher Edling. 2009. "Tocqueville and analytical sociology," in Mohamed Cherkaoui and Peter Hamilton (eds.), *Raymond Boudon. A Life in Sociology. Essays in Honour of Raymond Boudon*. Oxford: Bardwell Press.
Hedström, Peter and Richard Swedberg (eds.) 1998. *Social Mechanisms. An Analytical Approach to Social Theory*. Cambridge University Press.
Hedström, Peter and Lars Udéhn. 2009. "Analytical sociology and theories of the middle range," in Peter Hedström and Peter Bearman (eds.), *The Oxford Handbook of Analytical Sociology*. Oxford University Press, 25–47.
Hollis, Martin. 1994. *The Philosophy of Social Science. An Introduction*. Cambridge University Press.
Homans, George C. 1967. *The Nature of Social Science*. New York: Harcourt, Brace & World
1984. *Coming to My Senses. The Autobiography of a Sociologist*. New Brunswick and London: Transaction Books.
1987. "Behaviourism and after," in Anthony Giddens and Jonathan Turner (eds.), *Social Theory Today*. Palo Alto: Stanford University Press, 58–81.
Kitcher, Philip. 1989. "Explanatory unification and causal structure," in P. Kitcher and W. Salmon (eds.), *Scientific Explanation. Minnesota Studies in the Philosophy of Science*, vol. 13. Minneapolis: University of Minnesota Press, 410–505.

Little, Daniel. 1991. *Varieties of Social Explanation. An Introduction to the Philosophy of Social Science*. Boulder: Westview Press.

2009. "The heterogeneous social: new thinking about the foundations of the social sciences," in Mantzavinos Chrysostomos (ed.), *Philosophy of the Social Sciences. Philosophical Theory and Scientific Practice*. Cambridge University Press, 154–78.

Mantzavinos, Chrysostomos. 2005. *Naturalistic Hermeneutics*. Cambridge University Press.

Manzo, Gianluca. 2010. "Analytical sociology and its critics," *European Journal of Sociology* 51(1): 129–70.

Menger, Carl. 1996 [1883]. *Investigations into the Method of the Social Sciences*. Grove City: Libertarian Press.

Morgan S.L. and C. Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press.

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

Opp, Karl-Dieter. 2005. "Explanations by mechanisms in the social sciences. Problems, advantages, and alternatives," *Mind and Society* 4(2): 163–78.

Platts, Mark. 1979. *Ways of Meaning*. London: Routledge and Kegan Paul.

Psillos, Stathis. 2002. *Causation and Explanation*. Montreal and Kingston, Ithaca: McGill–Queen's University Press.

Raz, Joseph. 2000. "Explaining normativity: on rationality and the justification of reason," in Jonathan Dancy (ed.), *Normativity*. Oxford: Blackwell.

Salmon, Wesley C. 1998. *Causality and Explanation*. Oxford University Press.

Sawyer, R. Keith. 2005. *Social Emergence: Societies as Complex Systems*. New York: Cambridge University Press.

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

Sen, Amartya. 1977. "Rational fools: a critique of the behavioral foundations of economic theory," *Philosophy and Public Affairs*. 6(4): 317–44.

Smith, Michael. 1994. *The Moral Problem*. Oxford: Blackwell.

Stinchcombe, Arthur L. 1991. "The conditions of fruitfulness of theorizing about mechanisms in social sciences," *Philosophy of the Social Sciences* 21(3): 367–88. Reprinted in Aage B. Sørensen and Seymour Spilerman (eds.).

1993. *Social Theory and Social Policy: Essays in Honor of James S. Coleman*. Westport: Praeger, 23–41.

Udéhn, Lars. 2001. *Methodological Individualism: Background, History and Meaning*. London: Routledge.

Watkins, J.W.N. 1957. "Historical explanation in the social sciences," *British Journal for the Philosophy of Science* 8: 104–17. Reprinted in P. Gardiner (ed.) 1959. *Theories of History*. New York: The Free Press.

Wippler, R. 1978. "The structural-individualistic approach in Dutch sociology," *The Netherlands Journal of Sociology* 14: 135–55.

Woodward, James. 2003. *Making Things Happen. A Theory of Causal Explanation*. Oxford University Press.

Part I

Action and mechanisms

1 Ordinary rationality: the core of analytical sociology

Raymond Boudon

Ordinary rationality: the core of analytical sociology

The notion of *analytical sociology* was originally created as a label for a sociology resting on well-defined and realistic principles potentially applicable to the various types of phenomena with which sociology deals. Analytical sociologists were dissatisfied both with the kind of sociology that treats social actors as *irrational*, as conditioned and moved by social and cultural forces; and with so-called *Rational Choice Theory*, which treats them as *rational*, but only in a very restricted sense of the term. I would like to offer instead a notion of *Ordinary Rationality*. The consequent *Theory of Ordinary Rationality* is capable of resolving this persistent dissatisfaction, and is I believe well-suited to become the core principle, or grammar, of analytical sociology.

Modern social scientists generally endorse an instrumental conception of rationality. Most would agree with Herbert Simon's statement that "Reason is fully instrumental. It cannot tell us where to go; at best it can tell us how to get there" (1983, 7–8).[1] This leads to the dualistic view that people rationally choose the means they use to reach their goals, or be faithful to their values; while their goals, values and beliefs

[1] H. Simon follows here a definition of rationality dominant in the English-speaking world, dominant in philosophy no less than economics. For Bertrand Russell, "Reason has a perfectly clear and precise meaning. It signifies the choice of the right means to an end that you wish to achieve. It has nothing whatever to do with the choice of ends" (1954, viii). Any action is teleological: it aims at reaching some end and must choose appropriate means to that end. But this choice includes in many cases *beliefs*. These beliefs rest upon *theories*, and these theories rest in turn on *assumptions*. Russell's statement supposes that his notion of *right means* can be given a precise meaning. Hence *beliefs*, the *theories* on which beliefs rest, and the *assumptions* upon which theories rest must be *valid* if the *means* is to be considered *right*. *Valid* here means *true* if the belief bears on a representation of the world; and *fair*, *good*, *legitimate* in the case of *should-be* beliefs. In other words, the notion of rationality held by Russell and Simon introduces the narrow assumption that determining which means is right is always a trivial operation. Why, on the other hand, could not the *ends* endorsed by a subject be explained by the fact that they are inspired by *beliefs* grounded upon *theories* which are themselves grounded upon *assumptions*?

are imposed upon them by social, cultural, psychological or biological forces over which they have little control, and of which they might even be unaware. Many social scientists are very uncomfortable with this dualistic theory of social action. The anthropologist Robin Horton (1993) has (in my view, rightly) dubbed this idea that the goals, values and beliefs of people are merely a mechanical product of their socialization or environment a *sinister prejudice*.

Major classical and modern writers implicitly endorse what I consider to be a much more satisfactory theory of social action, elaborating a much broader view of rationality.

In the following, I will seek to:

1. make clear what I mean by *ordinary rationality*;
2. illustrate the usefulness of the *Theory of Ordinary Cognitive Rationality* or more concisely the *Theory of Ordinary Rationality* (TOR), using various examples; and
3. list the advantages of the theory from the scientific perspective.

The notion of ordinary rationality

Let us assume X represents some objective, opinion, or normative or positive belief. The *Theory of Ordinary Rationality* (TOR) assumes that a subject will probably endorse X if he has the more or less conscious impression:

1. that X is grounded on a system of reasons {S} including statements which appear to him to be individually acceptable and mutually compatible; and if he has the additional impression that
2. no alternative system of reasons {S'} is available which might be preferable to {S} and which would lead to an alternative objective, opinion, or normative or positive belief.

Scientific beliefs represent the most ready application of this theory. We accept Torricelli's assumption that quicksilver rises in an empty tube as a result of atmospheric pressure. The alternative Aristotelian explanation that quicksilver rises because nature abhors a vacuum is weaker, since it introduces an anthropomorphic notion. Moreover, if a barometer is carried to the top of a tower or a mountain, the indicated pressure falls. The Aristotelian assumption cannot explain this variation.

This example suggests three conclusions:

1. the belief that Torricelli's theory is true is here *rational* in the sense that it is *caused by reasons*;

2. the example illustrates Hollis' formula that rational action has the unique feature that it "is its own explanation" – the explanation of why the belief is endorsed is self-sufficient *because* it is rational; and
3. the rationality at work here is *instrumental* only in a trivial sense. It aids me in reaching my goal of choosing between two theories.

Beyond that, it provides me with a method which is *general* in the sense that I could apply it to any set of theories, and *universal* in the sense that anybody could appreciate and use it.

Hence what I call the Theory of Ordinary Cognitive Rationality, or more briefly the Theory of Ordinary Rationality (TOR),[2] postulates that this basic process is more or less at work whatever the nature of X – whether X is an opinion, a normative or representational belief, or an objective.

Before proceeding I would like to make some further remarks and draw some distinctions.

Deviations from the ideal-typical case

First remark. The case where a social actor endorses X because it is grounded on a system of reasons stronger than any other available system of reasons is *ideal-typical*.

Of course, there is nothing to say that it should always be possible to associate a set {S} of statements satisfying the two conditions of ordinary rationality to *any* X. It is impossible in many cases to decide whether a set of reasons {S} is better than {S′}, or whether the reverse is true.

To avoid another possible misunderstanding, ordinary thinking generally applies the ideal-typical model in an approximate fashion. As Pareto has argued, any belief is associated with reasons, but these reasons are often invalid; as in the case where a subject approves some X on the basis of a syllogism making X congruent with *natural* requirements, but where the major and minor statements employ the notion of *nature* with different meanings – as, according to Pareto, in the case of the belief that private property is *natural*.

The system of reasons mobilized by social actors can be invalid in many other ways. Thus, I can endorse a statement that *X is true* on the

[2] In other texts I have described this as *general* rather than *ordinary*. The two attributes are associated with important characteristics of the theory: the first insists on its logical character, the second on the continuity between ordinary and methodical thinking. The best description would be *judicatory*. It is used by Montaigne (*judicatoire*) to denote beliefs grounded on reasons, and by Max Scheler (*urteilsartig*) to describe Adam Smith's grounding of moral sentiments upon reasons. Unfortunately *judicatory*, *urteilsartig* or *judicatoire* are unfamiliar words.

basis of some particular reasons, but ignoring other reasons leading to the opposite conclusion, or because I am unable to imagine an alternative system of reasons. This situation can be observed in the context of ordinary deliberation, as shown by many experiments or observations made by social and cognitive psychologists; but it is also a property of scientific deliberation or discussion, as many studies of scientific controversies have shown.

Just as instrumental rationality can be perfect as an ideal, but is in reality in most cases *bounded*, ordinary cognitive rationality can also be unbounded as an ideal, and even in some actual cases. But it is usually *bounded*: for lack of access to relevant information, or because influenced by cognitive incompetence or of cognitive strategies,[3] or due to the interference of conflicting goals, or a particular *passion*.

Factual statements and principles

Second remark. The statements belonging to a set {S} grounding an objective, value or belief X normally belong to distinct categories. Some are *factual* statements, while others are *principles*. It is possible, at least in theory, to confront factual statements with the real world, whereas principles by definition cannot be proved. They are conjectural statements. The Torricelli example illustrates the two categories: it includes the *principle* that anthropomorphic statements – more precisely, final causes – should be excluded from scientific explanation of natural phenomena, together with the *factual* statement that air is weighty.

The history of science again provides a straightforward illustration of these ideas. Max Weber stated quite correctly that "Keine Wissenschaft is voraussetzunglos" ["There is no science without principles"] (1995 [1919]). This raises the question of whether the preference for a principle P1 rather than P2 can be considered rational.

The rationality of this preference in fact rests on the meta-principle according to which a principle P1 is preferable to a principle P2 if the theories inspired by P1 tend to be more satisfactory than the theories inspired by P2. As the universe of the theories inspired by a principle is open, it is never possible to consider a principle to be genuinely *proved*.

[3] As Tocqueville wrote, "we believe a million of things on the faith of others." Even the fact that suicide is more frequent in a network in which one of the members has committed suicide (Hedström) can be interpreted as the effect of cognitive strategies: I admired him/her; he/she had existential problems similar to mine; he/she showed me the "solution."

Table 1.1 *Types of system of reasons*

System of reasons	*Strong*	*Weak*
Context-free	Scientific beliefs (Torricelli)	Private property is *natural* (Pareto)
Context-dependent	Rain dances, miracles (Durkheim)	Beliefs of civil servants (Tocqueville)

Context-free and context-dependent beliefs

Third remark. Cognitive rationality can be *context-dependent* or *context-free*. A scientific belief aims at being context-free. Other beliefs are brought about for reasons that individuals belonging to one context consider certain, while those in another context do not.[4] Rain dances are, for example, considered by some societies to be effective, but not by modern Western societies.

Four ideal cases

The previous distinctions generate four ideal cases. The system of reasons grounding a belief in the mind of a social actor can be strong or weak, and context-dependent or not. It can be *context-free and strong*, as in the case of scientific beliefs. It can be *context-free and weak*, as in the above-mentioned example drawn from Pareto, where a belief is grounded upon reasons employing a notion with varied possible meanings. It can be *context-dependent and strong*, as in the case of the beliefs in miracles or in the effectiveness of rain dances, as shown by Durkheim. It can be *context-dependent and weak*, as when civil servants conclude from weak reasons that only state-ruled agencies can serve the interests of the public – as observed by Tocqueville.

The main thesis of the Theory of Ordinary Rationality

Taking these remarks into account, the main thesis of TOR is that *reasons* are the *causes* of X, where X can be a *normative* or *positive belief*, a *value*, a *means* or an *objective*.

[4] For the sake of clarity, I have abandoned the expression *subjective rationality* as used in my earlier writings. I had used it to characterize the case where ordinary cognitive rationality operates in one context, but not in other contexts – as in the case of rain dances. I now prefer to restrict this expression to cases where the reasons grounding a belief are related to personal idiosyncrasies.

In the following, using examples from various fields, I will suggest that TOR, once clarified, is the best candidate to form the core principle, or grammar, of the social sciences. I will consider successively examples where the Xs to be explained are *representational beliefs*, *normative beliefs*, *phenomena of consensus*, *long-term and medium-term evolutionary phenomena*, *personal objectives* and *practical solutions to classical interaction dilemmas*, such as the Prisoner's Dilemma. By introducing these examples I will suggest that TOR is a useful theoretical framework within which all kinds of phenomena of interest to the social sciences can be explained.

Representations as products of ordinary rationality

To begin with I will consider the case where X is a *representational belief*, and generally a belief which can be expressed by a statement of the type *X is true*.

I will not return to the issue of scientific beliefs. I will instead present here two examples involving representational beliefs generally regarded to be *irrational*.

Beliefs in miracles

Durkheim, and Renan before him, asked: why do people from earliest times up to the eighteenth century so easily believe in the existence of miracles? Why was this idea eradicated, if only partially? The response to this from Renan and Durkheim is: so long as the idea that natural phenomena are governed by *laws* has not been established, the contrast between unexplained phenomena and phenomena explained by *natural laws* was inconceivable and so had no meaning. *Miracles* is the name *modern* Westerners gave to events reported by the Holy Scriptures which today appear inexplicable in terms of natural law, but which then seemed normal to their contemporaries, if unexpected. The idea of *natural laws* slowly became well-established as physics, chemistry and then the life sciences made intensive use of it, so that events inexplicable in terms of natural laws came to be considered the product of illusion. All the same, the idea of a "miracle" was not entirely abolished, because the principle that any natural phenomenon will be the product of law-governed events cannot, as a principle, be proved. Moreover, contemporary philosophy of science treats science as aimed at the discovery of *mechanisms*, for example, the Darwinian evolutionary mechanism, rather than the discovery of *laws* explaining the natural phenomena. However, we do not yet know how to explain phenomena such as the

formation of the *eye* as the effect of explicit mechanisms. It is for this reason that Rousseau's argument that complex natural phenomena, just like complex cultural phenomena such as Homer's *Iliad*, cannot be considered to be the effect of chance mechanisms retains its appeal for some, as in the belief in *intelligent design*.

Peasants against monotheism

Why, asked Weber, were Roman civil servants and army officers attracted by monotheistic cults such as Mithraism imported from the Middle East, while Roman peasants felt deeply hostile to these cults and remained faithful to the traditional polytheistic Roman religion? Their hostility to Christianity was so deep-seated that the word *paganus* – peasant – was adopted by Christians to describe those who were hostile to Christianity: the heathens.

Let us concentrate on the hostility of peasants to Christianity. Weber argues that peasants had trouble accepting monotheism because the uncertainty characteristic of natural phenomena, an essential dimension of their everyday life, to them seemed incompatible with the idea that the order of things could be subjected to a single will: a notion implying a minimal degree of coherence and predictability. So, besides other more or less contingent factors, the main reason that peasants rejected monotheism was that Roman peasants, as good followers of Popper's falsification theory, had the distinct impression that monotheistic theory was incompatible with the data they were able to observe. This analysis also explains also why an impressively large number of *saints* appeared in the Christian world during the early centuries of our age. Thanks to the saints, Christianity was made more palatable to peasants, since it once more became polytheistic.

Axiological rationality

The idea of ordinary rationality led me to a further step: treating *axiological rationality* as a special form of ordinary cognitive rationality. Axiological rationality can be formally defined: given a system of statements {Q} containing at least one axiological statement and concluding that the axiological statement N is valid, all the components of the system of statements {Q} being acceptable and mutually compatible, it is axiologically rational to accept N if no alternative system of statements {Q}′ preferable to {Q} and implying a preference for N′ over N is available.

As in the case of representational beliefs,

1. this situation is *ideal-typical*;
2. it is difficult in many cases to decide whether N is preferable to N′;[5]
3. a system of reasons can be *flawed*; and
4. it can be *context-dependent or not*.

This definition formalizes the intuition contained in Weber's *axiological rationality*, but which had already been outlined by earlier writers. Adam Smith for instance explains in his *Wealth of Nations* the feelings of fairness or unfairness prompted by the wages of different occupations by making such feelings the effect of *strong* reasons, in the sense that people could not easily envisage more acceptable alternative systems of reasons.

Why, he asks (Smith 1976 [1776]: book 1, ch. 10), do we consider it normal that the public executioner is paid a relatively high salary? His qualification is minimal. His job supposes a low level of education and competence. He is – thank God – drastically underemployed. But since his job is the most disgusting of all a reasonably high wage should be paid as compensation. Some other wage differences rest upon more complex systems of reasons. Thus, Smith's contemporaries generally thought that coal miners should be paid more than soldiers. The two jobs require a low level of qualification. It takes a short time to train a miner and a soldier. Both are exposed to the risk of death. But the death of a miner is spontaneously interpreted as an *accident*, while the death of a soldier is regarded as a *sacrifice* for the sake of the country. Consequently, the soldier is entitled to symbolic rewards, while the miner is not. Since the two jobs are comparable from the viewpoint of qualification and exposure to risk, the principle *equal contribution, equal reward* requires that the miner receive a higher salary than the soldier, in compensation for the fact that, by contrast with the soldier, he is not entitled to glory and other symbolic goods. This statement can itself be considered as deriving from the principle of the *dignity of all*. Like any principle, it cannot be proved.[6]

[5] Today's theorists of axiological feelings are inclined to emphasize *dilemmas*, in contrast to classical writers – such as Kant, Smith, Weber or Durkheim – who are instead concerned with the issue of the source of *convictions*. An exclusive concern with dilemmas generates and/or expresses a relativistic view of axiological feelings. Michael Sandel's interest in dilemmas for example is related to his defence of communitarianism, which obviously rests upon a relativistic stance.

[6] This notion is taken into consideration by Kant, but in an abstract and static fashion. The elaboration which I here present under the label TOR, inspired in particular by Adam Smith, Emile Durkheim and Max Weber, has the advantage that it can explain a very large number of concrete axiological preferences.

To use a concept advanced by Smith in his *Theory of Moral Sentiments*, the relative consensus emerging over the question of whether one job should be more or less highly paid than another derives from the sets of reasons related to the presence of an *impartial spectator*,[7] involving the way that individuals seek to elaborate systems of reasons that stand a chance of acceptance by all.

In Smith's analysis, people react the way they do when they learn that a particular job is paid as it is for reasons which are only marginally of the *instrumental* type, but which are essentially *cognitive*.

Feelings of fairness

Many other examples in which ordinary cognitive rationality appears to support some feelings of fairness could be invoked, as also where empirical reciprocally observed feelings of fairness can be satisfactorily explained by TOR.

On the whole, the empirical and theoretical literature on the relationship between inequality and fairness shows that public feelings should be explained by the ordinary rationality approach. The conception of the *impartial spectator* clearly implies that there are valid reasons to consider some types of inequalities acceptable and in no respect unfair, while others are perceived to be illegitimate and unfair (Forsé and Parodi 2007). Thus:

1. Functional inequalities are not perceived by the *impartial spectator* to be inequitable. He easily recognizes that rewards should be indexed to aptitudes, responsibilities, competence and/or the contributions displayed by social actors.
2. They are not perceived to be unfair inequalities resulting from the aggregation of free choices or decisions on the part of a group of individuals. The pecuniary rewards for media celebrities or popular sportsmen and women are often thought to be *abnormally high*, but not *unfair* or inequitable, since such rewards are the outcome of the free individual choices made by their fans or supporters.
3. In principle, when the contributions of two individuals are of an identical value, they should be equally rewarded. But few people consider it to be unfair that two persons having the same job and doing the same task are unequally rewarded if they belong to unequally rich or dynamic firms or regions. Thus if plumber A is as competent

[7] The *impartial spectator* is a crucial notion in Smith's *Theory of Moral Sentiments*, and which remains implicit in his later *Wealth of Nations*.

as plumber B but is employed by a firm facing serious economic difficulties, it would be accepted that his salary was lower, even though his level of competence is the same.

4. Neither are inequalities between incommensurable activities considered unfair. It is possible to argue with Adam Smith that miners should be paid more than soldiers, but it would be difficult, for instance, to argue that climatologists should be paid more or less than the managers of supermarkets.
5. Inequalities the origin of which is unknown, or inequalities which cannot be defined as functional or dysfunctional – whether or not they reflect differences in competence or achievement – are not considered inequitable. This is an important point, for an overall income distribution is the product – to an unknown extent – of functional inequalities, of dysfunctional inequalities, and of inequalities of which it is impossible to say whether they are functional or not, that is to say, whether or not they reflect differences in competence, achievement or contribution. For this reason, and according to some illuminating but unfortunately isolated observations, people do not consider the reduction in range of the overall income distribution to be a major political objective. However, when global inequalities are strongly increasing, as in contemporary Western democracies, or when they lead to a *dual* society, as has become established in Brazil, negative feeling is aroused. This is very likely founded upon Rousseau's statement that social peace supposes that "the rich are not too rich and the poor not too poor."

By contrast, inequalities which can clearly be interpreted as privileges are considered deeply unfair – as when a business leader who has led his firm into decline, or even bankruptcy, is dismissed with a substantial payoff, or when a political leader uses his position to generate illegitimate advantage for his own benefit.

In sum, once observations made by the social sciences are synthesized they demonstrate that the public views inequality as fair or unfair with respect to ordinary rationality.

Consensus as a product of ordinary rationality

Ordinary rationality can also explain why certain institutions or states of affairs prompt social consensus, often after lengthy discussion and in many cases serious, violent or even bloody struggle. A few examples can be introduced as illustration.

Income tax

Over a long period democratic societies struggled with the question of whether, and in what form, an income tax should be introduced. Following lengthy political debate and conflict the idea was accepted, and income tax defined as proportional (*flat tax*). At a further stage, consensus emerged over the idea:

1. that the notion of an income tax was a good one;
2. that income tax should be progressive; and
3. that it should be moderately progressive.[8]

These three principles describe the currently prevailing situation in most Western democratic countries. They are widely accepted, with the exception of a few dissenting economists, since the impartial spectator has recognized that these three principles can be easily legitimated for reasons that are widely accepted.

These reasons are as follows. As Tocqueville has already described (1991 [1835]), modern societies are roughly composed of three social classes entertaining mutual relations of cooperation and conflict. The three classes are:

1. the rich, who have at their disposal a significant surplus which can be converted into political or social power;[9]
2. the middle class, which enjoys a more or less important surplus, but insufficient in size to convert it into political or social power;
3. the poor.

Social cohesion, social peace, the principle of the dignity of all, requires that the poor benefit from a subsidy, primarily from the middle class because of its numerical significance. However, the middle class would not assume its share if the rich were not to accept bearing the cost of subsidizing the poor to a greater extent than the middle class, and this conforms with elementary principles of justice. It can be concluded from this that income tax should be progressive. On the other hand, it must be moderately progressive, since the principle of efficiency would be violated if the tax was too brutally progressive, for the rich would have an incentive to transfer their resources abroad, generating a loss for the national community.

One can therefore legitimately conclude that the consensus on income tax results from a set of convincing reasons, accepted by common sense

[8] I borrow here from Ringen (2007a and 2007b).
[9] In the second volume of his *Démocratie en Amérique*. See Boudon (2005).

on account of their validity. Once sufficiently informed, any citizen belonging to any one of the social classes should accept the idea that a moderately progressive income tax is a good thing. The validity of this argument is responsible for the consensus, and its stability through time. Nonetheless, some citizens, swayed by their interests, prejudices or passions will very likely be hostile to the idea. There will always be dissenting economists. But they command no following, since they overlook the axiological dimension of the question. A too-brutally progressive income tax would violate the principle of efficiency. A flat tax would not be considered fair by the middle class.

This analysis explains other facts; for example, that differences between Scandinavian and continental European countries are diminishing over the years given the influence of the principle of efficiency.

Long-term evolution as a product of ordinary rationality

Ordinary rationality can also explain long-term evolutionary phenomena. Durkheim (1960 [1893]) observed that the evolution of judicial sentencing in Western societies is characterized by a secular trend toward more lenient sentences. Furthermore, an increasing number of acts having negative effects on other persons are dealt with by civil rather than penal law. An increasing number of cases are prosecuted in lower-level courts than hitherto.

This long-term evolutionary phenomenon derives mainly from a basic process: when a new form of sentence appears to be as equally effective in terms of dissuasion as a former one, and also better from the viewpoint of some particular criterion, the new type of sentence tends to be accepted and selected by the *impartial spectator*. In other words, a basic two-stage mechanism is at work in this type of evolutionary processes:

1. production of innovation;
2. rational selection of the innovation.

Durkheim has rightly maintained that social conditions can facilitate or thwart this basic mechanism, such as for instance increasing demographic density and the resulting increasing division of labor. This dependence on external conditions explains why, in this case as in many other cases, ordinary rationality is context-bound.

The foregoing analysis can easily be applied to our modern world. The death penalty tends to disappear from modern democratic societies notably because it has been repeatedly shown to have no dissuasive power. Moreover, it renders judicial error irreparable and is obviously

cruel. These reasons lead the *impartial spectator*, in the long term, to prefer other types of penalties, such as life sentences. The latter will therefore tend to be rationally selected.

Personal objectives as products of ordinary rationality

Personal objectives also can be explained within the frame of ordinary rationality, rather than, as with Gary Becker (1996), in the frame of instrumental rationality. Thus in my *Equality, Education and Opportunity* (Boudon 2006 [1973]), I introduced the idea that students:

1. tend to fix the social level or type of occupation they wish to achieve by taking as a reference point the status reached by people with whom they have a relationship, and
2. that they seek then to guess the probability of achieving the educational status capable of giving them a serious chance of obtaining this type of social status.

Once this system of reasons is modeled, it reproduces in a satisfactory fashion a statistical set of aggregate data. Moreover, the theory predicts which educational policy is better suited to the reduction of inequality of educational opportunity.[10] Several studies have been inspired by this ordinary rational approach to personal objectives.[11]

Benefits from the Theory of Ordinary Rationality

Some final remarks on the scientific benefits to be expected from moving from the current instrumental view of rationality to TOR can be introduced at this point.

Avoiding solipsism

TOR has the advantage of avoiding the objection of *solipsism*, while preserving the postulate of *methodological individualism* and the benefits of the rational theory of action deriving from the fact that "rational action is its own explanation" (Hollis 1977). The representations and evaluations of individuals rest upon reasons which they perceive as valid, and hence likely to be shared by others.

[10] The conclusions of Boudon (2006 [1973]) are confirmed in the case of modern Germany by Müller-Benedict (2007) using data from the Pisa study.
[11] Among the most recent and remarkable, see Manzo (2009).

Avoiding proceduralism

TOR avoids also at another level the weaknesses of procedural theories, such as those of Habermas or Rawls, which assume that representations and evaluations would be shared *if and only if* they can be considered to have been generated by acceptable procedures. In this formula, the *only if* is erroneous. Procedural rationality is much more popular today than substantive rationality. It nonetheless faces serious objections. Scientific representations are at least in principle generated by acceptable procedures. But, as Pareto sarcastically noted, the history of science is a graveyard of false ideas. It can be concluded that valid procedures cannot guarantee the validity of representational, nor of normative beliefs. They can only *facilitate* the generation of valid beliefs.

Solving RCT deadlocks

Needless to say, one of the important features of TOR is that it can easily solve many questions that Rational Choice Theory (RCT) is unable to solve. I will just mention a few examples.

As game theory rests upon the axioms of RCT, several situations of interaction appear to have no *solution*, in the sense that the theory is unable to recommend any satisfactory line of action to social actors. These are notably the classical structures of the *Prisoner's Dilemma*, the *insurance game*, the *chicken game* or the *battle of the sexes*. They have inspired a very extensive literature exactly because they have no *solution*, as long as one insists on strictly applying the RCT axiomatic. Leading social thinkers and sociologists have all understood that the only correct *solution* consists in taking steps to modify the structure, if feasible. Rousseau well understood in his *Discourse on the Origin of Inequalities* that the way to solve an *insurance game* is by breaking the structure through the introduction of a constraint. Olson has clearly seen that collective action can be made unlikely when social actors are trapped in an *n-person Prisoner's Dilemma* structure, and also that the solution is to undermine the structure by some innovation, such as the *closed shop* or the production of *selective incentives*. The famous *crossroads problem* has no solution within the RCT axiomatic, since RCT provides no way of choosing between the two Nash equilibria generated by the problem. The *solution* is again to introduce an innovation, in this case in the form of a priority rule. The interesting side of Axelrod's (1984) pages on the repeated Prisoner's Dilemma game is not that the *TIT for TAT strategy* provides a solution, but that, if actor A

cooperates in the first move, this sends to B the *signal* that, if he cooperates, A will probably go on cooperating in the following moves. This *expressive strategy* introduces an element that goes beyond the RCT axiomatic, if taken literally.

The TOR renders empty dispositional variables ineffective

Theories which, like RCT, rest upon an instrumental conception of rationality, cannot avoid introducing *dispositional variables* in order to explain the objectives, beliefs or values of social actors where these parameters of social action cannot be explained with the aid of instrumental rationality. Dispositional variables are however often tautological and ad hoc. They have generally a descriptive rather than explanatory function. Moreover, they imply the metaphysical and *sinister* view that men can be mechanically *conditioned* by social, cultural, psychological or biological forces of which they are unaware. Dispositional variables are on the whole the Achilles heel of the social sciences.

TOR proposes to substitute an explanation of beliefs in terms of context-bound rationality, rather than of *dispositions*. Two examples can be given.

1. In his analysis of magical beliefs, Weber (1968 [1920–21]) starts from the uncontroversial fact that *primitive peoples* do not know the law of the transformation of energy, from which he deduces that they have no *reason* to differentiate between the maker of rain and of fire, which Westerners consider to be obvious.
2. The Pharisees believed in the immortality of the soul, while the Sadducees did not. Why? Because they had since childhood been conditioned to these beliefs? Such an explanation explains nothing. The difference is in fact due to *context-bound reasons*. Weber explains that the Pharisees are for the most part shopkeepers, while the Sadducees are the reservoir from which the political elite is drawn. For the former, notions of exchange and of the equity of exchanges represent categories which they use in their everyday life. For this reason they are happy to learn that the soul is immortal, so that inequity represented by merits which remain unrecognized and sins which remain unpunished in this world will be corrected in the life after the death. The Sadducees have no reason to be attracted by an idea which appears to them opaque. Weber here does no more than to impute plausible *contextual reasons* to the two categories. His explanation has nothing to do with pseudo-explanations employing dispositional variables.

A general theory

Last but not least, TOR represents a genuinely general theory, in the sense that existing theories, as RCT or TBR, can be treated as special cases generated by introducing additional restrictions.

It should also be emphasized that I have been dealing with *social* actions, that is, with actions of interest to *sociologists*: individual actions the reception of which by other actors is taken into consideration by the actor. I have not been considering such actions from the standpoint of psychology, for from this point of view such actions are clearly not always rational. My claim that the social actor should be treated as rational, in the sense of TOR, is a *postulate for the social sciences*, not a philosophical statement describing the essence of men.

Answer to an objection

One objection is frequently raised against sociological theories that describe human behavior as cognitively rational: they are accused of *intellectualism*, of ignoring the role of emotion or violence in social relations. A profound remark by Voltaire (1957 [1763]) suffices to answer this objection:

> Those who sacrifice their blood and their life do not sacrifice in the same fashion what they call their reason. It is easier to lead one hundred thousand men into battle than to compel the mind of a true believer to submit.
>
> [ceux qui sacrifient leur sang et leur vie ne sacrifient pas de même ce qu'ils appellent leur raison. Il est plus facile de mener cent mille hommes au combat que de soumettre l'esprit d'un persuadé].[12]

Do not the most violent social conflicts turn on values and ideas?

REFERENCES

Axelrod, R. 1984. *The Evolution of Co-operation*. New York: Basic Books.

Becker, G. 1996. *Accounting for Tastes*. Cambridge, MA: Harvard University Press.

Boudon, R. 2005. *Tocqueville for Today*. Oxford: Bardwell.

2006 [1973]. *L'Inégalité des chances*. Paris: Hachette. *Equality, Education and Opportunity*. New York: Wiley.

Durkheim, E. 1960 [1893]. *De la division du travail social*. Paris, Presses Universitaires de France.

[12] Du protestantisme et de la guerre des Cévennes, *Œuvres historiques*, Gallimard, La Pléiade, pp. 1275–80.

Forsé, M. and M. Parodi. 2007. "Perception des inégalités économiques et sentiments de justice sociale," *Revue de l'OFCE* 102: 484–539.
Hollis, M. 1977. *Models of Man: Philosophical Thoughts on Social Action*. Cambridge University Press.
Horton, R. 1993. *Patterns of Thought in Africa and the West*. Cambridge University Press.
Manzo, G. 2009. *La spirale des inégalités*. Paris: Presses de l'Université de Paris-Sorbonne.
Müller-Benedict, V. 2007. "Wodurch kann die Soziale Ungleichheit des Schulerfolgs am stärksten verringert werden," *Kölner Zeitschrift für Soziologie* 59(4): 615–39.
Ringen, S. 2007a. *What Democracy is For. On Freedom and Moral Government*. Oxford and Princeton: Princeton University Press.
2007b. *The Liberal Vision*. Oxford: Bardwell.
Russell, B. 1954. *Human Society in Ethics and Politics*. London: Allen & Unwin.
Simon, H. 1983. *Reason in Human Affairs*. Palo Alto: Stanford University Press.
Smith, A. 1976 [1793]. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Oxford University Press.
Tocqueville, A. de. 1991 [1835]. *La démocratie en Amérique, in Tocqueville*. Paris: R. Laffont.
Voltaire 1957 [1763]. "Du protestantisme et de la guerre des Cévennes," *Œuvres historiques*. Gallimard, La Pléiade, 1275–80.
Weber, M. 1968 [1920–21]. *Gesammelte Aufsätze zur Religionssoziologie*. Munich: Mohr.
1995 [1919]. *Wissenschaft als Beruf*. Stuttgart: Reklam.

2 Indeterminacy of emotional mechanisms

Jon Elster

Introduction

Emotion-based actions involve causal mechanisms in three steps (see Figure 2.1). First, there is a belief (sometimes a perception) that triggers an emotion. Second, the emotion triggers an action tendency. Third, the action tendency may be strengthened, weakened or blocked by other motivational forces. I discuss these steps in the remaining sections of this chapter. The final section offers a brief conclusion.

Let me first explain how I use the term "mechanism." As in my recent work (Elster 1999: a, ch. I; Elster 2007: a, ch. 2) I define mechanisms as *frequently occurring and easily recognizable causal patterns that are triggered under generally unknown conditions or with indeterminate consequences*. In my conception of mechanisms, they are defined by the feature of *indeterminacy*. In what I call *type A* mechanisms, the indeterminacy concerns which of several possible mechanisms will be triggered, e.g. sour grapes versus forbidden fruits. In *type B* mechanisms, it concerns the net effect of several opposite mechanisms that are triggered simultaneously, such as the income effect and the substitution effect.

I shall not here try to argue for the usefulness of this conception, but simply point to the main implication for my purposes here, which is that the relations between belief and emotion and between emotion and action are not one–one, but one–many or many–one. The examples given later will clarify this idea. I should add that my idea of indeterminacy is a purely epistemic one: the mechanism or the net effect are *indeterminate as far as we know*, not in and of themselves, an idea that hardly make senses outside quantum mechanics. In some cases, progress of knowledge may enable us to identify the conditions under which this or that mechanism will be triggered or what the net effect will be if both are triggered.

As this article is largely a methodological reflection on, and extension of, some of my earlier writings on emotion (Elster 1999, 2009a), I do not give references for all the specific examples I cite.

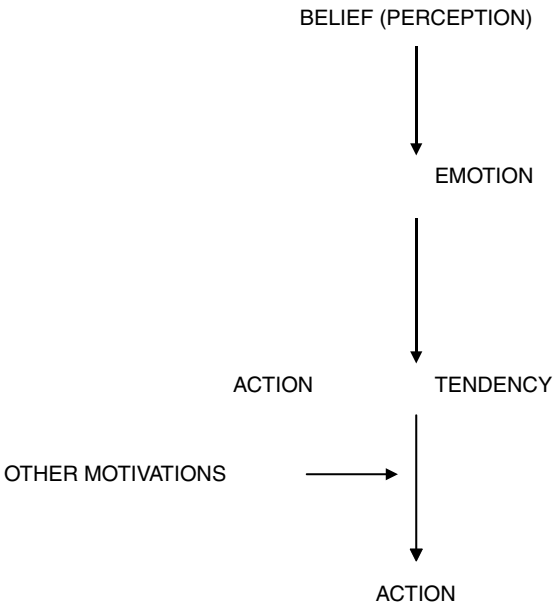


Figure 2.1 Belief, emotion and action.

For my purposes here, a careful definition of what counts as an emotion is not necessary. I emphasize mainly their cognitive antecedents and the action tendencies they generate, while neglecting other features. I shall only be dealing with core, garden-variety emotions, not with borderline cases. With the exception of love, they are mainly dark or negative: anger, indignation, fear, hatred, envy, ingratitude, shame and guilt.

Emotion and its antecedents

I shall focus on cases in which the antecedent of emotion is a *belief*, a propositional attitude in an epistemic mode ranging from (any degree of) subjective probability to certainty. Although mere possibility may also trigger emotion ("what if my child was killed in a car accident?"), I shall disregard this case since the emotion is less likely to trigger action. The same comment applies to emotions triggered by representations of action. Horror films usually do not cause people to flee the movie theater.

A more relevant exception is that of emotions being triggered by (non-propositional) *perceptions*. In a standard example, the mere sight of a snake-like shape on the path may cause fear through a pathway

that goes directly from the sensory apparatus to the amygdala, without passing through the cortex (LeDoux 1996). From the point of view of the social sciences, however, the more important effect of perception occurs when it is added to cognition rather than substituting for it. Abstract knowledge gains in vividness and motivating power when it is accompanied by visual cues. Some ex-smokers who are afraid of relapsing keep color photographs of smokers' lungs on the walls of their apartment. The sight of a beggar in the street may trigger more generosity than the knowledge of starving children in Africa. Propaganda often relies more on images than on words. Petersen (2005) illustrates this effect with a picture showing a graphic image of a happy and grinning Serb cutting the throat of a young Albanian boy. The caption urges the reader not to let Serbs return to Kosovo. The poster provided no new information, since everyone already believed that the Serbs committed atrocities; it only made the information more vivid and present.

Although emotions usually have a short half-life, they may persist for very long if the abstract belief is reinforced by constant visual reminders. Although we might think that killings would leave stronger memories than confiscation of property, the opposite can be the case. Referring to the twentieth-century descendants of those who had their property confiscated in the French Revolution, one historian wrote that "Generations forget more quickly spilled blood than stolen goods. By their continued presence under the eyes of those who had been despoiled of them, the fortunes providing from the national estates maintain an eternal resentment in the souls" (Gabory 1989: 1063). Table 2.1 shows a sample of important belief–emotion connections.

The third-party emotion that I refer to as "Cartesian indignation" was first identified by Descartes (1985: Art. 195), with the important proviso that if B *loves* C the indignation is transformed into anger (1985: Art. 201). Descartes was also the first to identify (in a letter to Princess Elisabeth of Bohemia from January 1646) what we might called "third-party gratitude," that is, B's positive feeling toward A caused by A's helpful behavior toward C.

Let me address the issue of indeterminacy with regard to some of these belief–emotion connections. Consider first the issue of *one–many relations*, that is, of a belief that can trigger one or more of several distinct reactions. I begin with some examples of type A indeterminacy.

As has often been noted (e.g. Tocqueville 2004: 60), the cognitive grounds for *envy* may also produce *emulation*. While the desire to do as well as a successful rival does not count as an emotion, the example shows that the grounds for envy are not sufficient conditions for that emotion to be produced. Similarly, the perception of danger may trigger

Table 2.1 *Belief–emotion connections*

Belief	Emotion
A imposed an unjust harm on B	B feels anger toward A
A imposed an unjust harm on C in the presence of B	B feels "Cartesian indignation" toward A
A is evil	B feels hatred toward A
A is weak or inferior	B feels contempt toward A
B feels contempt toward A	A feels shame
A has behaved unjustly or immorally	A feels guilt
A has something that B lacks and desires	B feels envy
A is faced with impending danger	A feels fear
???	B loves A
A suffers unmerited distress	B feels pity toward A
A has helped B	B feels gratitude toward A

either visceral (emotional) fear or prudential fear. Fleeing danger out of visceral fear can be irrational, if it causes us to go from the frying pan into the fire. A plausible example is provided by the estimated 350 excess deaths caused after 9/11 by Americans using their car instead of flying to wherever they were going (Gigerenzer 2004). By contrast, it appears that no excess deaths were caused by people switching from train to car after the attacks in Madrid on March 11, 2004. In this case, the reason may be that the Spanish were habituated to terror by decades of ETA actions, and had come to adopt an attitude of prudential rather than visceral fear (López-Rousseau 2005). In general, however, it may be hard to predict which of the two fear reactions will occur.

For some other one–many relations, consider cases involving three agents. In one case, A intentionally causes B to kill C. Will the relatives or neighbors of C feel anger toward A or toward B? Assume that A is a resistance movement in a German-occupied country, B is the occupational force and C is the as-yet uncommitted population. If the Italian resistance movement kills a German officer and the Germans respond by killing 100 civilians chosen at random, the anger of the population may be directed toward the resistance and not, as the resistance leaders may have hoped, toward the Germans (Cappelletto 2006). In another case, A offers a privilege to B but not to C. Will C feel anger toward A or envy toward B? Assume that A is the French king before 1789, B the French nobility and C the French bourgeoisie. Tocqueville argued that the bourgeoisie predominantly felt envy toward the nobility (see Elster 2009b: ch. 7 for references and discussion). In both cases, the

emotion experienced by C and/or the target of the emotion are subject to indeterminacy.

Further examples are provided by some of the "perverse" or "irrational" emotional reactions that moralists from Seneca to La Bruyère have delighted in exploring. When A unjustly harms B, the reaction of A may be anger rather than guilt: "Those whom they injure, they also hate" (Seneca, *De Ira*, II.23). According to La Bruyère (1885: 283), "The generality of men proceed from anger to insults; others act differently, for they first give offence and then grow angry." The explanation may lie in the pridefulness that prevents some persons from admitting that they have anything to feel guilty about: they seek a reason for having offended, and this reason then propels them to further offenses. Similarly, men "often hate those who have done them kindnesses" (La Rochefoucauld, Maxim 14) and in fact hate them *because of* the kindnesses (Seneca, *De Beneficiis*, III.1). Along similar lines, efforts by the rich to alleviate envy by generosity may in fact exacerbate it, by substituting an even more enviable property for the one that triggered the emotion in the first place.

I now turn to some examples of type B indeterminacy, in which the belief of an agent A about an agent B causes A to have two distinct and simultaneous emotions toward B. Thus I do not consider cases in which a belief might cause A to have two distinct and simultaneous emotions toward B and C, as would be the case if the French nobility felt both envy toward the nobility *and* hatred toward the king.

A first example involves *fear and anger*. According to Aristotle, "you cannot be afraid of a person and also at the same time angry with him" (*Rhetoric* 1380a). This does not seem right. Although in the typical case anger, unlike fear, is backward-oriented, Aristotle himself tells us that a person feels anger "because the other has done *or intended to do* something to him and his friend" (*Rhetoric* 1378a; my italics). Hence if I confront an enemy who intends to cause me some undeserved harm, the intention makes me afraid and its injustice makes me angry. Assuming that the action tendency of fear is to flee the enemy (see below, p. 58 for reservations), it might be overcome by the action tendency of anger, which is to fight him.

A second case involves *fear and hatred*. We may think of this as the "tyrant's dilemma" (see also Roemer 1985): draconian measures of repression may cause his subjects to fear him and to quell opposition, but also to hate him and to stimulate opposition. Hence Seneca advises kings to show mercy:

> Kings by clemency gain a security more assured, because repeated punishment, while it crushes the hatred of a few, stirs the hatred of all. The inclination

to vent one's rage should be less strong than the provocation for it; otherwise, just as trees that have been trimmed throw out again countless branches, and as many kinds of plants are cut back to make them grow thicker, so the cruelty of a king by removing his enemies increases their number; for the parents and children of those who have been killed, their relatives too and their friends, step into the place of each single victim. (*On Mercy*, I.8)

A third example involves the simultaneous feelings of *envy and sympathy* caused by the success of a friend. Hume (1960: 375–6) has an explicit statement of this duality:

In general we may observe, that in all kinds of comparison an object makes us always receive from another, to which it is compar'd, a sensation contrary to what arises from itself in its direct and immediate survey. A small object makes a great one appear still greater. A great object makes a little one appear less. Deformity of itself produces uneasiness; but makes us receive new pleasure by its contrast with a beautiful object, whose beauty is augmented by it; as on the other hand, beauty, which of itself produces pleasure, makes us receive a new pain by the contrast with any thing ugly, whose deformity it augments. The case, therefore, must be the same with happiness and misery. The direct survey of another's pleasure naturally gives us pleasure, and therefore produces pain when compar'd with our own. *His pain, consider'd in itself, is painful to us, but augments the idea of our own happiness, and gives us pleasure*. (my italics)

To be sure, a pain/pleasure duality is only one aspect of the emotions. From this passage, it is not clear whether Hume thought the dual emotions of sympathy and envy would also go together with dual action tendencies and, if so, which would dominate. Would my friend's success cause me to praise him or to denigrate him in the presence of third parties?

Consider now *many–one relations* between beliefs and emotions, that is, cases in which a given emotion can be triggered by qualitatively different beliefs. In the case of *anger*, the triggering condition is commonly defined as the belief that one has been unjustly treated. This is indeed a frequent cause of anger, but far from the only one. Seneca asks, "why do we see the people grow angry with gladiators, and so unjustly as to deem it an offence that they are not glad to die? They consider themselves affronted, and from mere spectators transform themselves into enemies, in looks, in gesture, and in violence" (*De Ira* II.2). Closer to us, which driver has never felt angry at pedestrians or bicyclists who impeded her progress? The mere belief that another person is an obstacle to the realization of my goal can trigger my anger.

With regard to *love*, the antecedents of the emotion are as multiple as they can be impenetrable. According to Stendhal (1980: 279) B's belief that A may – but also may not! – love B is a necessary condition for B's loving A. It is clearly not a sufficient condition. B may also believe that A has all sorts of wonderful qualities that makes him "lovable," but it

is often hard to tell whether the belief is an effect of the emotion or its cause. B's belief that A can offer her something she needs may also play a role. The belief that A *needs her* can fuel that belief, if B herself needs to be needed. In some cases, at least, B's love is triggered by B's *perception* of A as beautiful or gracious rather than by B's beliefs about A.

Finally, "irrational" guilt and shame can arise even when the agent does not believe she has done a bad thing or that others feel contempt for her. Thus the Holocaust created "survivor guilt" as well as "bystander guilt." People may also feel guilt knowing that they could have done something to prevent a bad outcome, even if at the time there was no way they could have known. (The same belief may trigger irrational anger in others.) Rape victims or disabled persons may feel shame even if they do not believe that others look at them with contempt.

Action tendencies and their emotional antecedents

If action is shaped by beliefs and preferences, emotions may influence action by their impact on either of these determinants. First, they may shape action by their impact on the *beliefs* that serve as premises for behavior. Thus the emotions of one jealous husband (M. de Rênal in *Le rouge et le noir*) make him believe, falsely, that his wife is faithful to him, while those of another (Othello) cause him to think, falsely, that she is unfaithful. Important as it is, I shall ignore this emotion–belief–action mechanism.

Second, emotions may shape action by their impact on *preferences* or, as I shall say, on the *action tendencies* that are the precursors of action (Frijda 1986). Each emotion is associated with a specific *action tendency*, which may also be seen a *temporary preference*. Here, I use the word "preference" in a somewhat large or loose sense. Strictly speaking, an agent can be said to have a preference for A over B only if both options are mentally present for him or her, and the agent, weighing them against each other, arrives at a preference for one of them. Consider, however, the following case. An agent knows that he is going to find himself in a dangerous situation and prefers, ahead of time, to stand fast in the face of the danger. This might, for instance, be an impending battle. Under enemy fire, however, he panics and flees. In this case, there may be no explicit comparisons of options, only an urge to run away that is so strong that everything else is blotted out. We may nevertheless think of the action as resulting from a preference for fleeing over fighting, reversing the earlier preference for fighting over fleeing. In actual cases, it will always be hard to tell whether the alternative of fighting makes a fleeting appearance on the mental screen.

Table 2.2 *Emotions and action tendencies*

Emotion	Action tendency
Anger, Cartesian indignation	Cause the object of the emotion to suffer
Hatred	Cause the object of the emotion to cease to exist or make it harmless by removing it
Contempt	Avoid the object of the emotion
Shame	"Sink through the floor"; run away; kill oneself
Guilt	Confess; make repairs; hurt oneself
Envy	Destroy or denigrate the envied object or its owner
Fear	Fight; flight; freeze; faint
Love	Approach, touch, help or please the object of the emotion
Pity	Console or alleviate the distress of the object of the emotion
Gratitude	Help the object of the emotion

In Table 2.2, the emotions that appeared as dependent variables in Table 2.1 appear as independent ones.

Consider first *one–many relations* between emotions and action tendencies. In the case of guilt, *confessing versus making repairs* are common, alternative, reactions. In an article, "La confession et la rédemption" [The confession and the redemption] in *Le Monde* of April 8, 2009, Sylvain Cypel draws a contrast between two traders, Michael Osinski and Stanford Kurland, who in different ways expressed guilt about their contributions to the recent financial debacle. The former chose confession. In an article "My Manhattan project" published in *New York Magazine* of March 29, 2009, he wrote that:

> Last month, my neighbor, a retired schoolteacher, offered to deliver my oysters into the city. He had lost half his savings, and his pension had been cut by 30 percent. The chain of events from my computer to this guy's pension is lengthy and intricate. But it's there, somewhere. Buried like a keel in the sand. If you dive deep enough, you'll see it. To know that a dozen years of diligent work somehow soured, and instead of benefiting society unhinged it, is humbling. I was never a player, a big swinger. I was behind the scenes, inside the boxes. My hard work, in its time and place, merited a reward, but it also contributed to what has become a massive, ever-expanding failure. For that, I must make a mea culpa.

By contrast, Kurland, a former president of Countrywide – America's largest mortgage lender until it crashed in 2008 – created a new firm,

Pennymac, devoted to buying distressed mortgages so cheaply that it can offer attractive repayment terms to struggling homeowners. Somewhat disingenuously, perhaps, as he also made a good profit on the transactions, Kurland presents this as "an act of redemption."

Similarly, *making repairs versus hurting oneself* may also be alternative action tendencies of guilt. Either tends to "restore the moral balance of the universe," the first by the wrongdoer undoing the loss of the victim and the second by imposing a comparable loss on himself. Thus if I feel guilty about evading my tax payments but know that the IRS will not accept anonymous checks, I might instead burn a suitable amount of money.

The best-known example is probably that of *fear*. Environmental stimuli can trigger one of four fear reactions: fight, flight, freeze or faint. Although the brain mechanisms that mediate freezing and fainting are quite different from those that trigger either flight or fight, little is known about when the former or the latter of the last two will be triggered. These claims, however, are related to research on animals, which act on the basis of perception rather than cognition. Casual evidence suggests that the fight–flight indeterminacy also applies to belief-triggered fear reactions in humans. Although it is sometimes claimed that fear induces fight only when flight is impossible, I do not know of any systematic evidence to that effect. In more complex social setting, "exit" and "voice" (Hirschman 1970) may perhaps be seen as analogues to flight and fight. To my knowledge, the determinants of the choice between exit and voice are poorly understood.

Finally, *hate* may either cause the desire to kill persons belonging to a hated category, or to make them harmless by removing them. Before they decided to carry out the Holocaust, the Nazi leaders contemplated sending the Jews to Madagascar. After the war, when many allied leaders wanted summary execution of the Nazi leadership, Henry Morgenthau proposed to deport the whole SS "out of Germany to some other part of the world" (US Senate 1967: 448). In criminal justice, the death penalty and a life sentence without parole may also be seen as alternative expressions of the same emotion.

Consider now *many–one relations* between emotions and action tendencies. The most obvious example is that both anger and fear can induce a tendency to fight. After 9/11, many American citizens and politicians strongly wanted to attack Afghanistan or Iraq. Some of these reactions may have been triggered by (prudential or visceral) fear, others by anger and the desire for revenge. Similarly, both anger and envy can induce a tendency to hurt another person with no benefit for oneself and at some cost. Consider the Ultimatum Game:

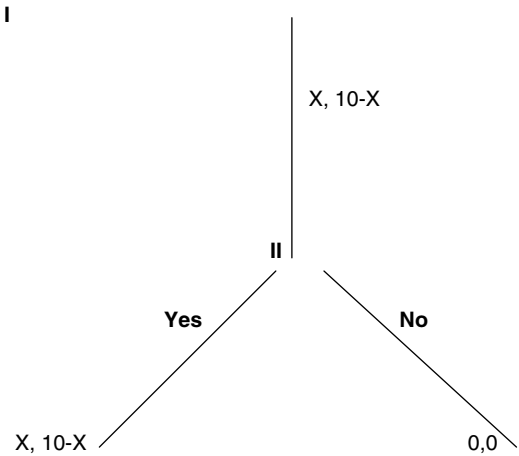


Figure 2.2 The Ultimatum Game.

In this game, one subject (the Proposer) offers a division of ten dollars between himself and another subject (the Responder). The latter can either accept it or reject it; in the latter case neither gets anything. A stylized finding is that Proposers typically offer around (six, four) and that Responders reject offers that would give them two or less (see Camerer 2003 for details and references). To explain this counter-interested behavior, we may assume that Responders will be emotionally motivated to reject low offers, and that self-interested Proposers, anticipating this effect, will make offers that are just generous enough to be accepted. The emotion in question might be either envy or anger. If it were envy, we should expect that the frequency of rejection of (eight, two) should be the same both when the Proposer is free to propose any allocation and when he is constrained – and known by the Responder to be constrained – to choose between (eight, two) and (two, eight). In experiments, the rejection rate is lower in the constrained case. This result suggests that Responder behavior is more strongly determined by the interaction-based emotion of *anger* than by the comparison-based emotion of envy. In other cases, envy – more specifically "black envy" – might be the more plausible explanation. Yet because of the unavowable nature of envy, it is often transmuted into anger (Elster 1999: 97–9, 350–3).

Finally, *envy and hatred* may induce similar action tendencies. When the envied object is a material possession or another's great beauty, the action tendency is to destroy it, using, for instance, fire or

vitriol as means. When A envies B's *character*, he can only destroy it by killing B.

From action tendencies to action

An action tendency is more than a mere disposition. It is an incipient action, or readiness for action, which can be identified by physiological changes and characteristic facial expressions. Whether it leads to an observable action depends on the presence or absence of other motivational factors.

In many cases, emotion-based action is blocked by *prudential motives*, such as (non-visceral) fear of loss or of punishment. In the Ultimatum Game, Responders may feel upset by a seven–three offer, and yet accept it out of self-interest. Even if they do not do so when stakes are low, they might accept an unfair offer if it amounts to a month's salary. In experiments, subjects are more willing to punish a non-cooperator when punishment is costless for the punisher than when it is costly (de Quervain *et al*. 2004). People who are motivated to a criminal action out of envy, anger or hatred may be deterred by the prospect of punishment. According to one study, "even in situations of supposed passion, people behave economically, weighing their actions' costs and benefits" (Shepherd 2004: 284). The visceral fear of a soldier before the enemy may be kept in check by the prudential fear of being shot "from behind" if he tries to flee. Emotions do not always make one "deaf and blind" to the consequences of action, although they sometimes have that effect.

On other occasions, an action tendency may be blocked by *social norms*. Because these are supported by feelings of *contempt* in observers of a norm-violation and of *shame* in the target of contempt, this blocking amounts to one emotion counteracting another. Soldiers are often kept from fleeing by the thought of what their fellow soldiers would say about them. In societies governed by codes of honor, many have preferred to risk their lives in duels or revenge to the "civic death" that would have followed a refusal to fight. In societies governed by the norm of turning the other cheek, some might refrain from taking revenge even when they could have done so at no tangible risk to themselves. If revenge is the spontaneous action tendency of anger it may, therefore, be either reinforced or weakened by social norms. By contrast, the impact of social norms on envy works only to *block* the action tendency. At least I do not know of any society in which one would be liable to blame and shame if one failed to act on an envious urge.

Finally, action tendencies may be blocked by the agent's *self-esteem*. If I have internalized the hierarchy of motivations that makes me

reluctant to act enviously in the presence of others, I may also abstain when I could have done so without risk of being detected. There may, in such cases, be an omission bias. A person who would not set his richer neighbor's house on fire might not call the fire department if he saw it burning.

Conclusion

Generally speaking, mechanisms are the basic building blocks – the nuts and bolts, cogs and wheels – of explanation in the social sciences. Emotional mechanisms form one set of such blocks, but there are many others. As an illustration, let me point to the importance of *psychic needs* in the explanation of mental states and thus, ultimately, of behavior. Here are a few examples:

- the need to have good and sufficient reasons for one's choices (Shafir *et al.* 1993);
- amour-propre: the egocentric need to preserve and promote a good self-image (La Rochefoucauld);
- the need for cognitive consonance (Festinger 1957);
- the need for cognitive equilibrium (Heider 1958);
- the need to believe that the world is just (Lerner 1980);
- the need for cognitive closure (Neurath 1913);
- the need for autonomy (Brehm and Brehm 1981);
- the need to act on motivations that occupy an elevated place in the hierarchy of motivations (Elster 1999).

I have already cited the last of these mechanisms as an explanation of why we might refrain to act on an envious urge. It seems likely that the other needs-based mechanisms might also interact, in various ways, with emotion-based mechanisms. It is of some interest to note that the theory of cognitive dissonance originated in an analysis of the need for a *justification of emotion*:

> The fact … which puzzled us was that following the [1934 Indian] earthquake, the vast majority of the rumors that were widely circulated predicted even worse disasters to come in the very near future. Certainly the belief that horrible disasters were about to occur is not a very pleasant belief, and we may ask ourselves why rumors that were "anxiety provoking" arose and were so widely accepted. Finally a possible answer occurred to us – an answer that held promise of having rather general application: perhaps these rumors predicting even worse disasters to come were not "anxiety provoking" at all but rather "anxiety justifying". That is, as a result of the earthquake these people were already frightened, and the rumors served the function of giving them something to be frightened about. (Festinger 1957: vi–vii)

According to this line of thought, we should be prepared to *reverse* the causal link from emotion to cognition that I explored earlier in the chapter (see pp. 51–6). This is not the occasion to pursue that issue. My purpose in citing the passage was only to emphasize that we should be prepared to see emotions within the larger framework of mental causation.

REFERENCES

Brehm, J. and S. Brehm. 1981. *Psychological Reactance: A Theory of Freedom and Control*. New York: Academic Press.

Camerer, C. 2003. *Behavioral Game Theory*. New York: Russell Sage.

Cappelletto, F. 2006. "Public memories and personal stories: recalling the Nazi-fascist massacres," in Francesca Cappelletto (ed.), *Memory and World War II*. Oxford: Berg.

Descartes, R. 1985. "Passions of the soul," in *The Philosophical Writings of Descartes*, vol. I. Cambridge University Press.

Elster, J. 1999. *Alchemies of the Mind*. Cambridge University Press

2007. *Explaining Social Behavior*. Cambridge University Press

2009a. "Emotions," in P. Bearman and P. Hedström (eds.), *Oxford Handbook of Analytical Sociology*. Oxford University Press.

2009b. *Alexis de Tocqueville: The First Social Scientist*. Cambridge University Press.

Festinger, L. 1957. *A Theory of Cognitive Dissonance*. Palo Alto: Stanford University Press.

Frijda, N. 1986. *The Emotions*. Cambridge University Press.

Gabory, A. 1989. *Les Guerres de Vendée*. Paris: Robert Laffont.

Gigerenzer, G. 2004. "Dread risk, September 11, and fatal traffic accidents," *Psychological Science* 15: 286–7.

Heider, F. 1958. *The Psychology of Interpersonal Relations*. New York: John Wiley.

Hirschman, A. 1970. *Exit, Voice and Loyalty*. Cambridge, MA: Harvard University Press.

Hume, D. 1960. *A Treatise on Human Nature*. Oxford University Press.

La Bruyère, J. de. 1885. *The Characters*. London: Nimmo.

LeDoux, J. 1996. *The Emotional Brain*. New York: Simon & Schuster.

Lerner, M. 1980. *The Belief in a Just World*. New York: Plenum Press.

López-Rousseau, A. 2005. "Avoiding the death risk of avoiding a dread risk. The aftermath of March 11 in Spain," *Psychological Science* 16: 426–8.

Neurath, O. 1913. "Die verirrten des Cartesius and das Auxiliarmotiv," translated in his *Philosophical Papers*, vol. I. Dordrecht: Reidel, 1983.

Petersen, R. 2005. "The strategic use of emotion in conflict: emotion and interest in the reconstruction of multiethnic states," unpublished manuscript, Department of Political Science, MIT.

Quervain, J.F. de, U. Fischbacher, V. Treyer, M. Schellhammer, U. Schnyder, A. Buck and E. Fehr. 2004. "The neural basis of altruistic punishment," *Science* 305: 1254–8.

Roemer, J. 1985. "Rationalizing revolutionary ideology," *Econometrica* 53: 85–108.
Shafir, E., Simonson, I. and Tversky, A. 1993. "Reason-based choice," *Cognition* 49: 11–36.
Shepherd, J. 2004. "Murders of passion, execution delays, and the deterrence of capital punishment," *Journal of Legal Studies* 33: 283–321.
Stendhal, B.H. 1980. *De l'amour*, ed V. Del Litto. Paris: Gallimard.
Tocqueville, A. de. 2004. *Democracy in America*. New York: American Library.
US Senate. 1967. *The Morgenthau Diaries (Germany)*. Washington, DC: US Government Printing Office.

3 A naturalistic ontology for mechanistic explanations in the social sciences

Dan Sperber

A naturalistic ontology for mechanistic explanations

There are several approaches in the social sciences that seek to provide causal explanations of social phenomena neither in terms of general causal laws nor in terms of case-specific narratives, but, at a middle level of generality, in terms of recurrent causal patterns or "mechanisms" (Hedström and Swedberg 1998). Typically, these approaches invoke micro-mechanisms to explain macro-social phenomena. Most of them, "analytical sociology" in particular (Hedström 2005), are versions or offshoots of methodological individualism. These individualistic approaches either stick to the "methodological" in "methodological individualism" and leave aside ontological issues, or else they are also individualistic in the metaphysical sense and deny the existence of supra-individual social phenomena that cannot be analyzed in terms of the aggregation of individual actions (see Ruben 1985).

The ontological challenge to which individualism responds is that presented by holistic approaches that place the social on a supra-individual level of reality. Another possible challenge, coming not from above but from below, that is, from the natural sciences, is generally not considered. The individuals invoked in individualism are not so much the individual organisms recognized in biology as the individual agents recognized in common-sense ontology. Individual agency is taken as a primitive in this approach, rather than as a tentative construct that should be unpacked and possibly questioned by psychology and biology.

Most mechanistic approaches, whether their individualism is just methodological or also metaphysical, show little interest in providing the social sciences with a naturalistic ontology, that is, one continuous with that of the natural science. The main goal of this chapter is to outline such a naturalistic ontology. But why should we want such an ontology in the first place? I don't, by the way, believe that the social sciences in general should systematically work within naturalistic ontology: many

of their goals, concern and programs are better pursued with the usual common-sense ontology. But when it comes to providing a scientific causal explanation of social phenomena, there are at least two reasons to prefer a naturalistic approach. The first reason is trivial: to the extent that it is possible, we would prefer our understanding of the world to be integrated, both for the sake of generality and for that of coherence.

The second, more interesting, reason to want a naturalistic ontology has to do with the quality of our causal explanations. Either the laws of physics admit of exception and social events provides such exceptions (and there is a Nobel Prize in physics to be won by doing sociology!), or else whatever has causal powers in the universe at large and among humans on earth in particular has them in virtue of its physical properties. Of course, this does not mean that social scientists should get involved in the physics of social causality. What it does mean though is that when we attribute causal powers to some social phenomena we should be able to describe it in such terms that its physical character is not a total mystery but raises a set of sensible questions that can be passed on to neighboring natural sciences, psychology, biology and ecology in particular, that directly or indirectly do ground their understanding of causal powers in physics.

The social sciences too, or at least scientific causal explanation programs in the social sciences, should ground their understanding of causal powers in physics, obviously in an indirect manner, by grounding it first in other natural sciences. Otherwise, we keep attributing causal powers to phenomena that we are not even able to locate in the time and space of genuine causal processes, and the chances are that we are making spurious causal attribution. At best, the correlations among events we describe might bear some more or less systematic relationship to actual causal processes but we are not in a position to ascertain this relationship, let alone to understand it. Of course, ontologically unconstrained causal-like descriptions may be good enough for one's purpose, but then one is not really aiming at scientific causal explanations of social phenomena. To put it in other terms, limiting oneself to a naturalistic ontology is a favorable – I am tempted to say, necessary – condition to arrive at sound causal claims.

In a tradition quite different from that of individualistic social science, drawing their inspiration from Darwinian evolutionary biology, there are other approaches aiming at providing scientific causal explanations of social, and more specifically cultural, phenomena. These approaches are not only mechanistic but also naturalistic; that is, they are committed to invoking only causal processes that can be

described in natural-science terms. However, I would argue, most of these approaches do not live up to their commitment.

The best known example of an evolutionary approach to culture is that of "memetics" inspired by the work of Richard Dawkins, and according to which culture is made of bits or "memes" that replicate themselves and propagate in a population through imitation (Blackmore 1999; Dawkins 1976). Just as biological evolution, cultural evolution is seen as largely governed by a process of Darwinian selection operating not among genes but among memes. Unlike holistic sociology, where people's behavior is largely determined by external forces larger than themselves, and individualism, where people are first and foremost agents determining their own behavior, memetics explains behavior, or at least cultural behavior, as determined by micro-forces operating within individuals and to a large extent controlling them, somewhat as viruses do.

Dual Inheritance Theory (Boyd and Richerson 1985; Cavalli-Sforza and Feldman 1981; Durham 1991) is another evolutionary approach to culture, and one that is more compatible with individualism than memetics. It describes people as collectively determining the evolution of culture by individually selecting cultural variants. In their selection, people are influenced by biologically inherited psychological biases that favor, for instance, imitating the majority, or the most prestigious individuals. According to the theory, mechanisms of cultural evolution differ in important respects from those of biological evolution, but the dynamics remain quite similar.

These evolutionary approaches to culture are innovative in many respects. They tend, however, to buy wholesale their catalog of cultural phenomena from the standard social sciences. Their naturalism consists to a large extent in providing natural causes – naturally selected psychological dispositions and ecological factors – for these non-natural social phenomena, or to adapt models of biological causality – more specifically of population genetics – to the cultural case. The cultural phenomena explained include however things such as religion, norms, art, racism, matrilineality, political hierarchy, and so on. Of course, these are postulated to have a proper naturalistic description, but nothing seriously approaching such a description is ever given.

Another mechanistic and naturalistic evolutionary approach to culture is the epidemiological approach that I have contributed to developing (Atran 1990, 2002; Bloch and Sperber 2002; Boyer 1994, 2001; Hirschfeld 1996; Sperber 1985, 1996, 1999, 2006), a hallmark of which is its insistence that a proper understanding of cultural phenomena and their propagation requires a deep understanding of the psychological

mechanisms involved (just as a proper understanding or standard epidemiological phenomena requires a deep understanding of individual pathology). The epidemiological approach takes seriously the ontological challenge of naturalism and suggests a way to provide a truly naturalistic ontology of the social. Here I outline how this can be done.

Relationships among neighboring disciplines may involve a difference of levels or a difference of scale. (The two terms, level and scale, are often used interchangeably, so I am, for expository purposes, sharpening a rather vague distinction.) To illustrate, contrast the case of psychology and neurology on the one hand and that of epidemiology and pathology on the other.

Until the cognitive revolution of the second half of the twentieth century, mental phenomena had no counterpart in the natural sciences. One could, of course, assert that mental phenomena occurred in the brain and postulate that they were wholly material, but there was no understanding whatsoever of how matter in general and brain tissues in particular might realize mental processes. The choice was then between pursuing a non-naturalistic psychology, and, as did behaviorists, pursuing a naturalistic psychology understood as a science not of the mind but of behavior. With the development of the mathematical theory of automata on the one hand, and of the neurosciences on the other, it is now possible to understand how matter in general and brain tissues in particular can process information. It is possible therefore to begin bridging gaps between psychology and biology. Psychological processes can be conceived as brain processes described in functional terms. At present, however, the concepts of psychology are not reducible to those of neurology, and it is contentious that they ever will be. So, the naturalization of psychology involves a matching in greater and greater detail of psychological and neurological descriptions but the two kinds of descriptions remain on quite distinct ontological levels. Neurology and psychology conceptualize different kinds of properties, and the concepts of one level cannot, at least for the time being, be defined in terms of the concepts of the other level.

Whereas the difference between neurology and psychology is one of levels, the difference between individual pathology and epidemiology is one of scale. Epidemiology studies the distribution of individual pathological conditions in a population. Epidemiology has its own concepts but not its own ontology. Its concepts are defined in terms of those of other disciplines: individual pathology, ecology, demography. Because it draws on several other disciplines, epidemiology is in a relationship of mutual relevance with all of them and of reduction with none of them. It is genuinely autonomous discipline, with strong bridges to other

disciplines, and without an autonomous ontology. Because the sciences from which it borrows its ontology are natural sciences, epidemiology is unproblematically a natural science too.

To better grasp the difference between differences of level and of scale, think of a zoom (a metaphor that was suggested to me by Bruno Latour, who however himself questions its appropriateness). You might initially be looking at, say, a single neuron. You might then zoom in to the scale of molecules, or you might zoom out to the scale of neuron assemblies, brain regions, or the whole brain, but at no point in this zooming-out do the objects you see become psychological rather than just neurological: you don't come to see thoughts or intentions with their contents. Even if, of course, what you are looking at includes the neurological realizers of such mental states, what you see is brain tissue all the way. Suppose by contrast that you are looking at an individual case of, say, the measles. You might zoom in within the organism to the scale of individual cells infected with the virus, or you might zoom out to the individual's environment, to her household, her community, the whole population to which she belongs. When zooming out from the individual to the population level the objects you see are still the same but they also become epidemiological ones. Whereas psychological objects are not neural objects seen at the scale of population of neurons – the difference is one of levels – epidemiological objects are just that: pathological objects seen at a the scale of a population in its environment – the difference is one of scale.

I propose a naturalization of the social science domain not on the model of psychology but on that of epidemiology. That is, I want to argue that social phenomena are patterns of psychological and ecological phenomena at a population scale. This project of naturalization of the social is made possible by the naturalization of the mental that is under way in the cognitive sciences. Let me explain how.

What makes a cognitive process cognitive is that it has as its function to secure a content relationship between its input and its output. In the case of perception, the input is a stimulus, the output is a mental representation, and the perception process aims at securing that the content of the mental representation should be a true identification of the input stimulus. In the case of memory processes, the input and the output are both mental representations and memory processes aimed at securing relevant content similarity between the two. In the case of inferential processes the input and the output are both mental representations and the inferential process aims at securing a relationship of justification: the content of the input, or premises, should justify that of the output, or conclusion. In the case of psycho-motor control, the

input is an intention, the output is a modification of the environment, and the psycho-motor control process aims at securing a relationship of realization: the modification of the environment should realize the intention. What the cognitive sciences are in the process of doing is explaining how material processes can reliably secure such relationships among representational contents (as in inference or memory) or between contents and the states of affairs they represent (as in perception and action).

Cognitive psychology studies the various processes that together make up the causal chains that go from the inputs to perception to the outputs of motor control. Actually, most cognitive psychologists study only one type of process involved in these causal chains: for instance perception, or memory, or inference. Among the criticisms that have been addressed to the standard cognitive psychology paradigm, many have to do with this relatively narrow view of cognitive processes. As critics have insisted, actual cognition is embodied, situated and distributed. I think that these criticisms are essentially correct, even if they are not as damaging as they are often claimed to be. Cognition is embodied: the brain is part of the body, and cognitive processes involve relationships not just between the environment and the central nervous system but also with the rest of the body (see, for example, Clark 1997). This is a truism, of course, the consequences of which are only now being systematically explored. Cognition is situated: the situations in which it occurs, in particular social situations, structure and guide cognitive processes (see, for example, Lave 1988). This is particularly obvious in the case of teaching and learning situations, but extends readily to all social situations and beyond. Cognition is distributed: many cognitive processes are realized not by a single individual but by a network that typically involves several individuals and artifacts (see, for example, Hutchins 1995). Is cognition in the brain, in the body, in the situation, in the network? In all of these and more. These descriptions should not be viewed as alternative theories but as complementary foci and scales.

The notion of a cognitive process should be understood so as not to be limited to processes located at an individual brain and its immediate periphery. Thinking of a cognitive process as a causal process that has the function of securing a content relationship (among representations or between representations and the state of affairs they represent) provides a simple and sensible way to broaden the picture. In particular, it justifies seeing all social processes as being also cognitive processes. Indeed, whatever else they do, social processes secure content relationships among the mental states and the actions of the people involved.

Cognitive processes are chained to one another, the output of some serving as input to others. This is true not only within individuals, but also across individuals. Communication, for instance, is a social cognitive process with two components, one of public expression on the part of the communicator, the other of interpretation on the part of the receiver. Interpretation, of course, takes as input the output of the expression process. The content relationship that the communication process aims at securing is a match in content between the communicator's meaning and the interpretation of the receiver. Communication itself is typically embedded in more complex processes that are both social – they involve interactions among individuals – and cognitive – they secure content relationships. Thus when one individual asks another to perform some action – be it an order or a request – the intentions of the first individual are satisfied through the action of another. When one individual gives testimony, her perceptions and inferences feed into the mind of others. When a group of people debate ideas or a course of action, they engage in a joint cognitive process that often none of them could have achieved on their own. All social interactions involve people acting on other people's minds. Conversely, too, any action on another person's mind is a social action.

Acting on other people's minds may not be the main feature or the main goal of social interaction. People may be after goods, space, food, sex, or whatever, but if their goals are social at all, they involve a cognitive dimension. When you buy an object, for instance, what you want is the object, but there is information transfer about the intentions of the buyer and of the seller, about the price, about the object. The price itself has an informational content the interpretation of which involves situating it in a historically extended causal chain that gives its value to the currency used. As another example, compare a person accidentally hitting another – not, or not yet a social interaction – and a person hitting another in an openly intentional way – unquestionably a social interaction. The difference that makes the second interaction social is that the hitter is not just transferring energy; he is also transferring information about his attitude to the victim, for instance about the grudge he may have or the rights he wants to enjoy.

A human group is criss-crossed by a complex flow of information. In this flow, not only information, but also people and things are being altered and moved around in an endless variety of ways. Still, let me insist: you can have social interactions without goods being made and handled, without bodies being moved, without sex, or without food, but you cannot have a social interaction without transmission of

information, and whenever information is being transmitted you have a social interaction. Let me also add, much of this flow of information goes through individuals and groups without being intentionally or even consciously transmitted. We transmit information through behaviors that do not have such transmission as their goal. Even when we engage in intentional communication, the information we actually transmit consists typically in both less and more than we intended. Most cognitive processes are unconscious, and so-called conscious processes are only partly conscious. Similarly, much of human behavior is unintentional, and many aspects of intentional behavior are not controlled by intentions.

I see no a priori reasons to give pride of place to intentions among the cognitive determinants of behavior. I am not, in other terms, indirectly arguing for yet another cognitive version of methodological individualism. When Peter Hedström writes: "Since changes in … social properties must be either intended or unintended outcomes of individuals' actions – how else could they possibly be brought about – they should be analyzed as such" (Hedström 2005: 5), I fail to see how the conclusion follows from the premises. Such outcomes of actions must also be expected or unexpected, planned or unplanned, emotionally loaded or not, and so on, but it does not follow that they should be analyzed as such (which is not to deny that it may, in some cases, be relevant to do so). The social cognitive causal chains I am talking about are typically infra-individual, or, if you prefer, sub-personal (a notion introduced by Dennett 1969), and, at the same time, trans-individual. They involve a variety of mental mechanisms, the formation and carrying-out of intentions being one or, more plausibly, a subset of them, an important one, no doubt, but not obviously more important than, for instance, the mechanisms of attention or of memory.

Standard epidemiological phenomena are causal chains of pathological events inside organisms and of events in the environment of organisms. Similarly, I claim, social phenomena are causal chains of mental events inside people and events in their common environment. These environmental events comprise behaviors and effects of behaviors such as transformation and movement of objects and people. Mental events can be better described in terms of a naturalized psychology. Environmental events can be described in plain materialistic terms, drawing on the appropriate natural sciences when relevant. So the nodes and links in social cognitive causal chains can be characterized naturalistically. The naturalistic challenge is to reconceptualize the social just on the basis of these causal chains.

Objection: two materially identical environmental events, say, a wink and a mock wink, may be quite different social events (simplifying an example of Gilbert Ryle made famous by Clifford Geertz; see Geertz 1973; Ryle 1971). The typical winker is aiming at secret communication with just one other person. The mock winker is drawing the attention of third parties on the failure of the winker to keep her communication secret. A description in terms of their mere material properties cannot account for the socially crucial difference between a wink and a mock wink. An eye twitch is an eye twitch. True, the twitch in a mock wink tends to be exaggerated, but it need not be. Is then the only way to go to replace the superficial material description with a 'thick' interpretative description, as Geertz argues? The alternative I am suggesting is to stick to the material description of the public events and to explain the difference by the fact that the wink and the mock wink occur at different places in social cognitive causal chains, and in particular have different mental causes and effects. To interpret an eye twitch as a wink is to attribute to it one kind of mental cause in one kind of chain of events; to interpret it as a mock wink is to attribute to it another causal history.

These two approaches – the interpretative and the naturalistic – don't differ in their intuitive understanding of what is happening. They differ in their ontology of meaning. For the interpretativist, meaning is public, it is in the public event, and is therefore beyond the reach of a naturalistic approach. For the naturalist, meaning is in the causal relationships of the public event to other events, in particular mental ones.

Just as cognitive events owe many of their properties to the fact that they are embodied, situated and distributed, social events owe many of their properties to the fact that they are mentalized, situated and distributed. Or, in terms I prefer, human cognitive and social events are what they are because they are embedded in social cognitive causal chains, and the chains involved in individual cognition and in social interaction are the same. They are just considered at different scales.

Even if you grant me that we might naturalize the difference between a wink and a mock wink, you may sensibly feel that the challenge of reconceptualizing the social just in terms of social cognitive causal chains is an excessively difficult or even impossible one. You might, more importantly, object that the project of reconceptualizing the whole domain of the social sciences in naturalistic terms, even if feasible, would be counterproductive since there is a wealth of accumulated knowledge and competence formulated in the current conceptual

framework that might be lost in the process. However, as the example of the wink suggest, the kind of naturalistic concepts of the social I am advocating are both significantly different but also closely related to more standard conceptualizations.

Let me give some very simple examples of the kind of reconceptualization I am advocating. Take a folktale such as Little Red Riding Hood. You can think of it as a collective representation that has evolved over time in European societies, has been taken up by a literary tradition aimed at a different social class from Charles Perrault and the Grimm brothers onwards, that expresses cultural attitudes toward unmarried women as preys, and so on. Naturalists ask: where in space, where in time is Little Red Riding Hood? Where, when and how does it enter into causal processes? And their answer is: the tale of Little Red Riding Hood is an abstraction, useful as such, but not to be confused with something with causal powers. What you have rather, in the world of causes and effects, is a social cognitive causal chain that extends over countries and centuries, that is made of public tellings and mental rememberings of indefinitely many versions of Little Red Riding Hood, millions and millions of micro events inside and among people. Causal forces apply at the level of these micro events and processes. Most stories told never reach a cultural level of distribution. Few culturally stabilized stories are as resilient as Little Red Riding Hood. One of the questions to ask then is what stable or variable properties of individual minds, of inter-individual encounters, and of the local environments where these occur explain the resilience of the tale – that is, the fact that its many versions stay close to one another – and also its evolution.

Take prestige, or to be a bit more concrete, take the intellectual prestige of Professor Jones. Prestige is characterized both by a content and by a distribution. The content has two aspects: an outstandingly positive evaluation of the intellectual merit of Jones, and the representation of this evaluation as being widely accepted. It is not important for prestige that the positive evaluation be justified – prestige need not be deserved – but it is essential that this evaluation be widely distributed and represented as such – prestige must be recognized to be prestige. To explain the prestige of Jones is to explain the joint distribution of two representations: that Jones is an outstanding professor and that a great many people agree that he is outstanding. To develop this explanation, one must identify the social and cognitive factors that, at every micro step, secure this distribution and stabilize its contents.

Examples such as tales or prestige illustrate, if anything, too easily the notion of a social cognitive causal chain. After all, the causal chains involved consist in an alternation of public and mental representations of similar content. The cognitive dimension is hardly contentious. But what about, for instance, an institution? An institution need not be mentally represented, and, in the case of complex institutions, it does not even lend itself to being mentally represented in an individual mind. Institutions are the paradigmatic examples of social things that seem irreducibly social. Of course, once we accept that cognition can be, and often is, distributed among people and artifacts, we have no difficulty in recognizing that institutions, however complex, involve such distributed cognition. But can we *characterize* institutions in terms of social cognitive causal chains?

Here is a proposal: institutions are characterized by an articulation of hierarchically related causal chains. At higher levels, they are distributed representations that prescribe how lower level representations (and behaviors, and artifacts) should be distributed, and the distribution of these higher level representations plays a causal role in the distribution of the lower level items. Take a folktale distributed by an extended social cognitive causal chain, what I would call a cultural cognitive causal chain. Add to it an extended distribution of a higher level representation with a normative content that prescribes, say, that this folktale is to be told on Christmas Eve. The distribution of this higher level representation indeed causes the tale to be told on Christmas Eve. Now instead of having, so to speak, a free-floating folktale, we have an elementary institution: a Christmas tale. More complex institutions, universities, churches, armies, markets, for instance, involve the articulation of many more social cognitive causal chains with a much greater variety of changes in the environment, but the principle is the same.

Most standard concepts in the social sciences are generalized and regimented versions of concepts deployed by social agents themselves. The social science concepts of status, class, caste, law, rights, contract, politics, state, religion, ritual, marriage, war, art and so on are borrowed and adapted from folk sociology. Social agents and social scientists alike attribute causal powers to the social phenomena denoted by these concepts. A marriage, say, is described as causing changes in rights and duties. From a naturalistic point of view, these are misattributions of causal powers. However, these attributions are generally not wide of the mark. On the contrary, there is a fairly systematic closeness between standardly misattributed causal powers and genuine causal processes.

The effects that social agents and social scientists attribute to, say, a marriage closely correspond to the effects of causal chains that distribute representations of this marriage. The effects that social agents and social scientists attribute to, say, a law correspond to the effects of causal chains that distribute representations of this law, and so on.

The naturalist social scientist reconceptualizes the social in terms of causal chains that distribute representations. The idea is not at all, let me insist, that mental representations are the social, or cause the social. Social cognitive causal chains contain events not just in minds but also in the environments, and the relative causal weight of mental and environmental events varies with the type of social phenomenon: mental events are relatively more important in literature as a social phenomenon, and environmental events are relatively more important in war as a social phenomenon. But the thread that links all social events is a cognitive thread that goes through minds, through the environment, through minds, through the environment, and so on, securing content relationships along the way.

So, to conclude, here is what I have argued:

- A naturalization of the domain of the social sciences is made possible by the ongoing naturalization of psychology.
- The ontology of a naturalized social science is a composite ontology, articulating naturalistic description of mental and environmental events.
- Precisely because, on this view, naturalized social sciences borrow the ingredients of their ontology from several different disciplines, their concepts and theories cannot be reduced to the concepts or theories of any one of these disciplines.
- The way in which naturalized social sciences renounce ontological autonomy secures their theoretical autonomy. In other terms, I am arguing for an ontological reduction without theoretical reduction.

Explaining social phenomena, in this perspective, is identifying the recurrent causal patterns or causal mechanisms that produce regularities in social cognitive causal chains. These regularities permit in turn to identify types of social phenomena (in a "population thinking" way, that is, without ever essentializing them; see Mayr 1970). Many of the type so indentified are likely to have close counterparts in folk and scholarly sociology, but they have a different ontology, one that comes with sound methodological constraints. Such constraints should be welcome when the goal is scientific causal explanation.[1]

[1] This chapter expands and revises an earlier text published in French (Sperber 2007).

REFERENCES

Atran, Scott. 1990. *Cognitive Foundations of Natural History: Towards an Anthropology of Science*. Cambridge University Press.

2002. *In Gods We Trust: the Evolutionary Landscape of Religion*. Oxford University Press

Blackmore, Susan J. 1999. *The Meme Machine*. Oxford University Press.

Bloch, Maurice and Dan Sperber. 2002. "Kinship and evolved psychological dispositions: the Mother's Brother controversy reconsidered," *Current Anthropology* 43(4): 723–48.

Boyd, Robert and Peter J. Richerson. 1985. *Culture and the Evolutionary Process*. University of Chicago Press.

Boyer, Pascal. 1994. *The Naturalness of Religious Ideas: a Cognitive Theory of Religion*. Berkeley: University of California Press.

2001. *Religion Explained: the Evolutionary Origins of Religious Thought*. New York: Basic Books.

Cavalli-Sforza, Luigi L. and M.W. Feldman. 1981. *Cultural Transmission and Evolution: a Quantitative Approach*. Princeton University Press.

Clark, Andy. 1997. *Being There: Putting Brain, Body and World Together Again*. Cambridge, MA: MIT Press

Dawkins, R. 1976. *The Selfish Gene*. Oxford University Press.

Dennett, Daniel. 1969. *Content and Consciousness*. London: Routledge and Kegan Paul.

Durham, W.H. 1991. *Coevolution: Genes, Culture and Human Diversity*. Palo Alto: Stanford University Press.

Geertz, Clifford. 1973. *The Interpretation of Cultures*. New York: Basic Books.

Hedström, Peter. 2005. *Dissecting the Social. On the Principles of Analytical Sociology*. Cambridge University Press.

Hedström, Peter and Richard Swedberg (eds.) 1998. *Social Mechanisms: an Analytic Approach to Social Theory*. Cambridge University Press.

Hirschfeld, L.A. 1996. *Race in the Making: Cognition, Culture, and the Child's Construction of Human Kinds*. Cambridge, MA: MIT Press.

Hutchins, E. 1995. *Cognition in the Wild*. Cambridge, MA: MIT Press.

Lave, Jean. 1988. *Cognition in Practice: Mind, Mathematics and Culture in Everyday Life*. New York: Cambridge University Press.

Mayr, Ernst. 1970. *Populations, Species and Evolution*. Cambridge, MA: Harvard University Press.

Ruben, David-Hillel. 1985. *The Metaphysics of the Social World*. London: Routledge & Kegan Paul.

Ryle, Gilbert. 1971. "The thinking of thoughts: what is 'Le Penseur' doing?" in *Collected Papers*, vol. II. London: Hutchinson, 480–90.

Sperber, Dan. 1985. "Anthropology and psychology: towards an epidemiology of representations (The Malinowski Memorial Lecture 1984)," *Man* (N.S.) 20: 73–89.

1996. *Explaining Culture: a Naturalistic Approach*. Oxford: Blackwell.

1999. "Conceptual tools for a natural science of society and culture, Radcliffe-Brown Lecture in Social Anthropology 1999," *Proceedings of the British Academy* (2001) 111: 297–317.

2006. "Why a deep understanding of cultural evolution is incompatible with shallow psychology," in N.J. Enfield and Stephen Levinson (eds.), *Roots of Human Sociality*. Oxford: Berg, 441–9.
2007. "Rudiments d'un programme naturaliste," in Michel Wieworka (ed.), *Les Sciences sociales en mutation*. Auxerre: Editions Sciences Humaines, 257–64.

4 Conversation as mechanism: emergence in creative groups

Keith Sawyer

Most explanations are causal explanations: "I ate a snack because I was hungry." This is a very simple causal explanation: the event of "being hungry" caused the event of "eating a snack." However, this statement, connecting two states at two successive moments in time, does not provide the full causal story connecting these two events. How, exactly, did "being hungry" cause "eating a snack"? A mechanist would argue that the complete explanation would have to describe the underlying mental state of "being hungry" and how this mental state then caused the neurons in the brain to make the body move into the kitchen and prepare the snack. This would be a very complicated story involving millions of neurons, as well as the relations between the stomach, the brain and the body.

This simple example demonstrates a general property of mechanistic explanations: they commonly describe processes in very complex systems that underlie what, on the surface, appears to be a simple causal relationship. Advocates of mechanistic explanation are most likely to be found in the philosophy of biology and in the philosophy of the social sciences, because both biology and the social sciences are centrally concerned with complex systems. A mechanistic explanation of an event (getting a snack) traces the causal processes leading up to that event by describing the components of a complex system and their interactions (the brain, the stomach sensors that detect hunger). They are *generative* explanations; they explain how the operations of the system generate the observed phenomena.

Many mechanists go beyond this definition to make two additional claims that I reject in this chapter. First, many mechanists define mechanistic explanation in opposition to deductive-nomological (DN) or covering law approaches. The DN approach explains an event by identifying initial conditions and a series of lawful generalizations that show how the event can be deduced from the initial conditions. An example of a covering law that accounts for my simple example, how to explain the act of eating a snack, would be the lawful generalization "When a

person is hungry, they are likely to eat" combined with the initial condition, "I am hungry," thus allowing the observer to deduce that I will prepare some food. I argue that the opposition of mechanistic explanation to the DN approach has been over-emphasized and is not necessary to the definition of mechanism. After all, even after the complete neurological account of hunger and snack-preparing has been developed, the simple regularity "hunger causes eating" still holds true (and is no doubt more useful in everyday life than the elaborately detailed account involving millions of neurons). Here I recommend Pierre Demeulenare's chapter in this volume, Chapter 9, which explores the notion of regularities versus special cases and argues that contra Peter Hedström, mechanistic and covering-law explanations are compatible; and in fact, they are more interrelated than most of us believe.

Second, many mechanists make the additional claim that a mechanistic explanation must be a lower-level one; a higher-level phenomenon is said to be explained when the lower-level system that gave rise to the phenomenon is sufficiently described: the components of the system, and their interactions. In this view, "hunger" is a mental state that must be explained in terms of the neuronal configuration of the brain that corresponds to that state. This form of mechanism is equivalent to methodological individualism in the social sciences, and to physicalist reductionism in the philosophy of science more generally. To perhaps take my hunger example a bit too far, imagine the twenty-person kitchen staff in a three-star restaurant; the explanation of how the kitchen works to generate hundreds of multi-course meals in one evening is a mechanistic account in terms of individual actions and interactions. The components of that system are individuals, just as the components of the brain are neurons. In both cases, the mechanistic explanation works at a lower level of analysis, describing system components and their interactions.

In this volume, Sperber's accounts of mental states (Chapter 3), and Elster's account of emotions (Chapter 2), are mechanistic explanations, and yet they do not refer to neurons. They present accounts of psychological mechanisms, not neurological mechanisms. By analogy, just as psychological mechanistic explanations can be constructed that do not have neurons as their components, social mechanist explanations can be constructed that do not have individuals as their components. Thus, contrary to most social mechanists, I argue that social mechanism is not definitionally identical with methodological individualism (just as psychological mechanism is not identical to neuroscience). Mechanisms exist at many levels of analysis. A sociological explanation could be a causally mechanist explanation even if it does not concern properties of

individuals. One could provide mechanistic explanations of large-scale social systems in which the components are smaller-scale social units (for Auguste Comte, it was the family, for example).

In my empirical studies of creative groups, including jazz and theater ensembles, I explain their performances by providing mechanistic accounts in which communicative interactions are the fundamental unit of the system. This approach is grounded in a long tradition of sociological study of interaction – including symbolic interactionism, ethnomethodology and conversation analysis – that goes back to at least Georg Simmel. In empirical practice, a focus on interaction has led to a focus on social mechanisms; social psychologists in general have focused on the interactional mechanisms that give rise to the emergence of group properties. The key mechanisms in these groups involve symbolic communication.

Those social mechanists who propose reductionist and individualist accounts of emergent social phenomena generally incorporate only simplistic accounts of human communication. In this chapter, I argue that social mechanistic explanations are strengthened when they incorporate the sophisticated symbolic aspects of communication. However, once symbolic communication is incorporated into a social mechanistic explanation, it becomes exceedingly difficult to formulate that explanation in terms of methodological individualism.

Improvisational encounters

A successful improvised performance is an emergent phenomenon, one that can be observed as it unfolds on stage. After years observing these creative groups – jazz and improvisational theater ensembles – I developed a strong intuition that the whole was greater than the sum of the parts. For example, the performance that emerges is not predictable, even knowing a great deal about the members of the ensemble. The performance that emerges cannot be reduced to the mental states and intentions of the individual performers. And the performance that emerges is novel – a unique creation. These three features – unpredictability, irreducibility and novelty – are classic characteristics of emergent phenomena (Sawyer 2005).

In an improvisational theater performance, how does the interactional frame emerge – the characters, motivations, relationships, and plot events and sequence? No single participant creates the frame; it emerges from the give and take of conversation. The frame is constructed turn by turn; one person proposes a new development for the frame, and others respond by modifying or embellishing that proposal.

Each new proposal for a development in the frame is the creative inspiration of one person, but that proposal does not become a part of the frame until it is evaluated by the others. In the subsequent flow of dialog, the group collaborates to determine whether to accept the proposal, how to weave that proposal into the frame that has already been established, and then how to further elaborate on it.

Example 1 is the first few seconds of dialog from a scene that the actors knew would last about five minutes. The audience was asked to suggest a proverb, and the suggestion given was "Don't look a gift horse in the mouth."

Example 1. Lights up. Dave is at stage right, Ellen at stage left. Dave begins gesturing to his right, talking to himself.

1	Dave	All the little glass figurines in my menagerie, The store of my dreams. Hundreds of thousands everywhere!	Turns around to admire.
2	Ellen		Slowly walks toward Dave.
3	Dave		Turns and notices Ellen.
		Yes, can I help you?	
4	Ellen	Um, I'm looking for uh, uh, a present.	Ellen is looking down like a child, with her fingers in her mouth.
5	Dave	A gift?	
6	Ellen	Yeah.	
7	Dave	I have a little donkey.	Dave mimes the action of handing Ellen a donkey from the shelf.
8	Ellen	Ah, that's – I was looking for something a little bit bigger …	
9	Dave	Oh.	Returns item to shelf.
10	Ellen	It's for my Dad.	

By turn 10, elements of the frame are starting to emerge. We know that Dave is a storekeeper, and Ellen is a young girl. We know that Ellen is buying a present for her Dad and, because she is so young, probably needs help from the storekeeper. These dramatic elements have emerged from the creative contributions of both actors. Although each turn's incremental contributions to the frame can be identified, none of these turns fully determines the subsequent dialog, and the emergent dramatic frame is not chosen, intended or imposed by either of

the actors. (Also note that the dialog is not obviously derived from the audience's suggestion, either, although the actors will later integrate it with the emerging frame.)

The emergence of the frame cannot be reduced to actor's intentions in individual turns, because in many cases an actor cannot know the meaning of her own turn until the other actors have responded. In turn 2, when Ellen walks toward Dave, her action has many potential meanings; for example, she could be a co-worker, arriving late to work. Her action does not carry the meaning "A customer entering the store" until after Dave's query in turn 3. In improvised dialogs, many actions do not receive their full meaning until after the act has occurred; the complete meaning of a turn is dependent on the flow of the subsequent dialog. This sort of retrospective interpretation is quite common in improvised dialog, and it is one reason that the emergent frame is analytically irreducible to the intentions or actions of participants in individual turns of dialog.

Improvisational encounters are different from the social phenomena that are usually studied by social mechanists. Social mechanists typically focus on middle-range phenomena, such as unemployment in Stockholm (Hedström 2005). The most immediate contrast is that the number of agents is quite small: in Example 1, there are only two agents. In contrast, many mechanistic explanations – particularly those using the computational simulation technique known as agent-based modeling – have hundreds or thousands of autonomous agents (e.g. Epstein 2006; Hedström 2005). There are three additional differences that lead me to argue that even though these social systems have as few as two agents, they nonetheless demonstrate the usual properties of emergence – including unpredictability, irreducibility and novelty. These three are:

1. a broad range of possible actions;
2. retroactive interpretation;
3. downward causation.

A broad range of possible action

At each point in the improvisation, an actor can choose from a wide range of moves to propel the dramatic frame forward. This unpredictability results in dialog that, at each turn, has a combinatorial complexity: a large number of next turns is possible, and each one of those turns could result in the subsequent flow of the dialog going in a radically different direction. This results in expanding combinatorics such as those in a chess game. Such moment-to-moment combinatorics often result in analytically irreducible phenomena.

Retroactive interpretation

It is often impossible to determine the meaning of a turn at the moment that the turn occurs. In Example 1, when Ellen walks toward Dave, her action has several potential meanings, and these ambiguities are not resolved until the subsequent turn of dialog. Before the subsequent flow of the interaction attributes a meaning to it, each of the two actors may have had a different interpretation of what the walking meant; it is not until later that these conflicting potential interpretations are resolved. This unfolding intersubjectivity results in collaboratively emergent phenomena that cannot be reduced to an analysis of the component individuals.

Downward causation

An individualist account of Example 1 would analyze each turn in terms of the causal influence of the prior turn of dialog, combined with the intentions and mental models held by the speaker, without any appeal to an analytically distinct social level of analysis. But with improvised dialogs, it is difficult to account for an individual's actions without making explicit reference to the independent causal force exerted by the emergent interactional frame.

The collaborative emergence of frames has been studied by several researchers in interactional sociolinguistics and conversation analysis, including Deborah Tannen, Alessandro Duranti and Charles Goodwin (Duranti and Goodwin 1992; Tannen 1993). These researchers have generally avoided arguing that the emergent frame is a real social phenomenon with autonomous social properties (Sawyer 2003b). Rather, the frame is considered to exist only to the extent that it is "demonstrably relevant" (Schegloff 1992) to participants – a classic interpretivist stance, one that rejects social realism.

In contrast, I argue that in improvised dialogs, the emergent interactional frame must analytically be considered to have its own causal force. Most sociologists generally believe that the causal impact of macro-social forces must be mediated through an individual's perception of them. However, in improvised dialog, people act without conscious awareness or reflection. If causal relations of constraint and enablement can be empirically identified in improvised dialogs, this would be evidence in support of the more general sociological claim that social emergents can have causal effects on individuals. I have identified such relations in several extended analyses of long improvised dialogs (Sawyer 2003b).

Improvisation as explained by methodological individualism

An individualist would hold that the interactional frame is nothing more than representations of it in the minds of individual participants; there is no higher level of analysis at which the frame emerges. The methodological individualist would explain improvised dialogs in terms of the mental representations of the frame held by the participants. The frame is not an autonomous social reality and it has no causal power.

Hedström and Swedberg's (1998) introductory chapter explicitly noted that "the principle of methodological individualism is intimately linked with the core idea of the mechanism approach" (p. 12), one of their four principles was "the general reductionist strategy in science" (p. 25), and both their examples and their typology of social mechanisms were methodologically individualist. They closed by implicitly indicating their opposition to social realism: "there exist no such things as 'macro-level mechanisms'." Thus social mechanists implicitly and explicitly deny a form of social realism that holds that emergent social properties can have downward causal powers (as advocated by the emergentists Archer 1995 and Bhaskar 1975/1997).

I hold that laws are compatible with mechanisms; mechanisms "explain" laws (Beed and Beed 2000; Bunge 2004; Elster 1998). However, it is possible that social laws may exist that are difficult to explain by reduction to micro mechanisms; if so, the scope of methodological individualism would be limited. In an earlier work (Sawyer 2005) I provided an account of emergence that I called *nonreductive individualism (NRI)* that shows how this could be so. I argue that some emergent social properties may be real, and may have autonomous causal powers, just like real properties at any other level of analysis. To the extent that social properties are real, social mechanist explanation may be limited to the explanation of individual cases that do not generalize widely, resulting in a case study approach rather than a science of generalizable laws and theories.

In developing NRI, I was heavily influenced by arguments surrounding the mind–brain relation; and many philosophers believe that these arguments can be generalized to apply to any hierarchically ordered sets of properties (Fodor 1989; Humphreys 1997: 3; Jackson and Pettit 1992: 107; Kincaid 1997: 76; Yablo 1992: 247, n. 5). NRI holds to a form of *property dualism* in which social properties may be irreducible to individual properties, even though social entities consist of nothing more than mechanisms composed of individuals.

The argument uses notions of supervenience, multiple realizability, and wild disjunction to argue that the relation between emergent social

properties and their realizing mechanisms is one of token identity, but not of type identity. In other words, on any token occasion, a social property must be realized by a mechanism involving its component individuals. However, on different occasions the same social property might be realized by different mechanisms (multiple realizability), and those different mechanisms might not be similar in any sociologically meaningful way (wild disjunction).

It is possible that some social properties that participate in causal laws are multiply realized in wildly disjunctive systemic mechanisms. If so, that social property is real and it has autonomous causal power (Sawyer 2003c). Thus like Fodor (1974) in the philosophy of mind, and Kincaid (1997) in the philosophy of social science, in such cases I argue that causal explanation need not cite underlying mechanisms. For example, the social property "competitive team sport" can participate in causal laws, and can play an important role in sociological explanation. Such explanations are quite standard in science more generally; each science standardly holds that its types and properties are real and that its properties participate in causal laws. Pressure laws hold regardless of what particles make up the gas, and regardless of the details of their relations and interactions. Airfoil laws hold with cloth sails and steel airplane wings, in varying temperature and humidity.

The methodological individualist's usual response to multiple realizability arguments is to claim that in such cases the many realizations of a given social property are lawfully and meaningfully related, so that multiple realization is a trivial and ignorable detail (e.g. Macdonald and Pettit 1981; Pettit, personal communication, 2002). A methodological individualist would have to argue that if a given social property is realized on different occasions by different mechanisms, those different mechanisms would nonetheless be quite similar and describable in basically the same explanatory framework. The methodological individualist might argue that on all occasions when a group has the property "competitive team sport," the individual members have similar intentional states (such as "belief") and the members are organized according to similar interpersonal bonds (such as "solidarity").

In the case of improvised dialogs, it is particularly difficult to make this argument.

An expanded conception of social mechanism

In improvised dialogs, there is always a continuing dialectic: social emergence, where individuals are co-creating and co-maintaining emergent interactional frames; and downward causation from those emergent frames. During conversational encounters, interactional

frames emerge, and these are collective social facts that can be characterized independently of an individual's interpretations of them. Once a frame has emerged, it constrains the possibilities for action. Although the frame is created by participating individuals through their collective action, it is analytically independent of those individuals, and it has causal power over those individuals. I refer to this process as *collaborative emergence* (Sawyer 2003a) to distinguish it from models of emergence that fail to adequately theorize interactional processes and emergence mechanisms. To account for these processes, sociologists must develop an account of the mechanisms of collaborative emergence that lead to these emergents.

When examining the extensive theoretical literature on emergence, in both philosophy and sociology, one finds a great deal of confusion, differences in terminology, and disagreement about the exact nature of emergent phenomena. For example, a key question remains unresolved: can emergent phenomena have autonomous downward causal power over the participating individuals? In improvisational theater, my empirical observations seemed to demonstrate "downward causation" – the emergent dramatic frame had constraining and enabling effects on the individual actors. I identified downward causal effects of emergent phenomena, through detailed interaction analyses of specific theater performances (Sawyer 2003b). As a consequence of these empirical observations, I developed the above theory of nonreductive individualism.

Emergence theories have existed long before the recent turn to mechanistic theories. Nonetheless, emergence and mechanism are quite compatible – in that both approaches emphasize processes and interactions among components in complex systems. Mechanistic approaches began to emerge in the 1990s, both in the philosophy of biology (Bechtel 2001; Bechtel and Richardson 1993; Craver 2001, 2002; Glennan 1996; Machamer *et al*. 2000) and in the philosophy of the social sciences (Elster 1989; Hedström and Swedberg 1998; Little 1998; Stinchcombe 1991).

Most social mechanists reject covering-law explanations of social processes and phenomena (cf. Markovsky 1997; Turner 1993). Rather than explanation in terms of laws and regularities, mechanists argue that sociology should provide explanations by postulating the processes constituted by the operation of mechanisms that generate the observed regularities. Such explanation is provided by the specification of often unobservable causal mechanisms, and the identification of the processes in which they are embedded (Hedström and Swedberg 1998). As such, mechanistic approaches implicitly assume a realist perspective

and reject empiricism (Aronson *et al*. 1995; Bhaskar 1975/1997; Layder 1990).

I believe that sociology has much to gain by a turn toward mechanism. In particular, I believe that an adequate account of emergence requires a mechanistic explanation. And as an example, I argue that mechanistic approaches are required to adequately explain improvised dialogs. I furthermore argue that these creative conversational encounters are at the core of all socially emergent phenomena, and that a mechanistic account of creative conversation is a necessary component of any account of social life. However, the appropriate mechanistic account would not be strictly individualist; it would incorporate individuals, symbolic interactions and emergent properties of the interactional frame.

All mechanistic accounts are necessarily reductionist in the sense that they reject "strong emergence" – a form of emergence in which emergent properties have autonomous causal power over the components of the system. Only within a covering-law framework can higher-level properties have causal power over lower-level properties. There can be no mechanistic downward causation within a single token system; downward causation can only exist as a lawful generalization across multiple cases (which are multiply realized). Yet, I argue that mechanists go too far in claiming that there can be no strong emergence; it is true that it cannot exist within mechanistic explanation, but it can hold for covering-law explanations.

Conversations are the mechanisms of social emergence

Sociologists who study emergence rarely incorporate the study of symbolic interaction; and sociologists who study symbolic interaction rarely consider emergence. Our goal should be to develop a theoretical framework that incorporates both symbolic interaction and emergence. Toward that goal, I begin this final section by briefly summarizing how these topics have been explained by two prominent paradigms in sociological theory over the past century: the structure paradigm, and the interaction paradigm (this is a summary of the more extensive account that appears in Sawyer 2005). The structure paradigm focuses on the relations between two distinct levels of analysis: the *individual* and the *social*. Most social mechanist accounts are of this type, in that they explore relations between two distinct levels of analysis, the individual and the social. The interaction paradigm is distinguished by its focus on an additional, intermediate level of interaction, symbolic action in

communicative episodes. I argue that neither of these two theoretical strands can adequately account for conversation and emergence. The solution is to take a mechanistic approach, one that can unify the best elements of both the structure and the interaction paradigms.

The structure paradigm

Social mechanists generally work within a long tradition in sociology that I call the structure paradigm (Sawyer 2005). The defining feature of this paradigm is its focus on the relations between two levels of analysis, the individual and the social. Traditionally, sociologists within the structure paradigm have fallen into two opposed camps: collectivists and individualists. Collectivists emphasize the ontological autonomy of the social level, and its causal power over individuals. Individualists argue that collective phenomena are epiphenomenal, and that all social forces must operate through the perceptions of them by individuals. What unifies these opposed camps is that both assume that conversation is epiphenomenal – it has no causal consequences, either for emergent macro phenomena or for individuals. Instead, the ultimate causal forces in social life are either institutions, networks and group properties (for the collectivist), or rational actions taken in the context of pairwise game-like encounters (for the individualist).

More recently, a third camp has emerged that blends elements of both; I call these hybrid theories. The most advanced hybrid theories are of the micro–macro link (e.g. Alexander *et al.* 1987; Knorr-Cetina and Cicourel 1981) and of the structure–agency link (Archer 1995; Giddens 1984). Although a theory of emergence must be a theory that relates the individual and the collective levels, even these advanced versions fail to explain emergence, because they have inadequate theories of the interactional mechanisms that connect the two levels of analysis. Ultimately, the structure paradigm cannot explain social emergence, because it does not incorporate theories of process, mechanism and interaction.

The interaction paradigm

The structure paradigm fails to account for emergence because it does not theorize the interactions and processes within a system's mechanisms. Thus it does not theorize the critical mediating link between the individual and the social levels of analysis: interactions between individuals. In the 1960s and 1970s, important alternatives to the structure paradigm began to emerge; I group them together using the term *interaction paradigm*. The interaction paradigm represented a fundamental

break; it focused on the processes and mechanisms of interaction neglected by the structure paradigm. Although this was a necessary step forward, ultimately I conclude that the interaction paradigm fails to account for emergence, although for very different reasons than the structure paradigm.

The defining feature of the interaction paradigm is the belief that properties of interaction are not derivable from the individual actions or agency of the members of the group, nor can they be derived from social structure. Interaction is an ontologically distinct level of analysis. There are fundamental properties and laws of interaction – based, for example, on semiotics, cybernetics or communication theory – that are not reducible either to individual properties or structural characteristics. Because interaction is not reducible to individuals or to structure, it is an autonomous level of analysis. The claim that interaction is a level of analysis with ontological status implies that it possesses properties that participate in causal relations, with causal effects on both structure and individuals.

Interaction is at the center of an old sociological tradition associated with Simmel, Cooley, Mead and the Chicago School of symbolic interactionism. Beginning in the 1960s and 1970s, symbolic communication has become central to a remarkably wide range of theories. These theorists have taken different positions on the relationship between interaction, individual and structure; what unifies them is that, in contrast to structure paradigm theorists – who believed that interaction was epiphenomenal – they give symbolic communication a prominent role. Versions of the interaction paradigm that became influential in the United States include symbolic interactionism, ethnomethodology and conversation analysis. Interaction paradigm theories in Europe include Bourdieu's notion of *habitus* (1972/1977), Foucault's discussions of *discourse* (1969/1972), and Habermas' theory of *communicative action* (1987). In both Europe and the United States, the interaction paradigm emerged at about the same time, and in both cases as a response to the inherent tensions of the structure paradigm.

The interaction paradigm denies that the social world has an objective, irreducible structure that constrains individuals in interaction. Macro-level concepts such as social structure and culture are considered to be abstractions that do not really exist nor have any causal force over individuals. Social reality can only be ascribed to concrete interactional processes and it can only be studied in terms of participating individuals' interpretations of it (Blumer 1962: 190). Macrostructural forces never operate directly on individuals, but are mediated through their interpretation by those individuals.

The interaction paradigm was a necessary intellectual development because it enabled researchers to undertake the close empirical study of the interactional processes of social life. These interactional processes had not been studied by the structure paradigm; a static synchronic focus led it to neglect the dynamic contingency of situated discourse. However, in making this antithetical move, the interaction paradigm was left with a problematic orientation toward emergent frames and structures. The interaction paradigm is uncomfortable with the idea that social phenomena might be irreducibly emergent. For example, the interactional frame remains under-theorized because a theory of the frame is necessarily partially collectivist and partially structuralist, and as such any theory of the frame seems, to an interactionist, to suffer from the same problems as the structure paradigm (Sawyer 2005). This stance has made it difficult for the interaction paradigm to study several aspects of social emergence, including how individual participants are constrained by macro-social forces extending far beyond the encounter, and how individual actions collectively result in the emergence of macro-social phenomena.

The interaction paradigm commonly leads to what I call *interactional reductionism*, the position that sociologists need only to study interaction: "macrophenomena are made up of aggregations and repetitions of many similar microevents" (Collins 1981: 988). Most of the symbolic interactionists and conversation analysts were interactional reductionists; they focus primarily on small-group encounters and do not directly address the emergence of the social from complex systems of individuals.

Although interactional reductionism is the dominant form of the interaction paradigm, there are a few hybrid theorists who have attempted to analyze both interaction and structure and their mutual relations (although without much impact on mainstream sociology). The concept of *discourse*, originating in 1970s French discourse analysis, has been central to many hybrid interaction theories (Sawyer 2002a). During the 1970s, French discourse analysts such as Paul Henry and Michel Pêcheux developed Althusser's concept of ideology into a theory of discourse. Pêcheux (1982) explored the relations that discourses have with ideologies; discourses, like ideologies, develop out of clashes with one another, and there is always a political dimension to writing and speech. Pêcheux explored the relationship between discursive formations and *ideological formations*; for Pêcheux, the discursive formation was the key concept that provided the causal link between social structures and individual consciousness: "Individuals are 'interpellated' as speaking-subjects (as subjects of *their* discourse)

by the discursive formations which represent 'in language' the ideological formations that correspond to them" (1982: 112). Discourse is a level of analysis that intermediates structure and individual. However, Althusserian discourse theory did not explore how structures emerge from interaction; rather, it emphasized that interaction is determined by structure.

Beginning in the 1990s, an effort known as *critical discourse analysis* emerged from British cultural studies, building on these 1970s notions of discourse and ideology (Fairclough 1995). Critical discourse analysis shares with Althusserian discourse analysis the attempt to simultaneously study both interaction and social structure, considering each as autonomous levels of social reality. However, in practice, critical discourse analysis has rarely documented or explained specific cases in which structural phenomena emerge from interaction. In most cases, the empirical studies of critical discourse analysis demonstrate how social structures reproduce themselves through interaction. These studies often demonstrate the mechanisms of class reproduction – a traditional Marxian concern – in spite of (or with the unwitting participation of) the creative agency of individuals. Critical discourse analysis has not proposed a theory of social emergence.

The lack of a theory of social emergence is a consequence of the failure to theorize a mechanism between social structure and interaction. The interaction paradigm has not explained the mechanisms whereby macro-social phenomena causally influence interaction. And, more importantly, the interaction paradigm does not explain how interaction could causally influence social structure; these influences would require a theory of social emergence, focused on explaining how interactional processes result in the emergence of social properties. Without a well-developed account of how interactional processes result in social emergence, interaction paradigm theorists have been unable to convince other sociologists that conversation is central to sociological theory.

- The interaction paradigm rejects the necessity of examining individuals (the realm of psychology) and macro-social structures (the realm of macro-sociology). Thus it seems that no concept of emergence is necessary.
- The interaction paradigm lacks ontological depth; its conception of reality has a narrow scope around the interaction level, and neglects the causally autonomous properties of both structures and individuals.
- The interaction paradigm has not theorized social constraint; in fact, interactional reductionists reject the existence of such constraint.

In sum, the interaction paradigm has no theory of social emergence – no explanation of how stable structures emerge from the joint collective actions of individuals engaged in social interaction. Accounting for causal relations between structure, interaction and individual requires a theory of the mechanisms of conversation that give rise to social emergence. A social mechanist approach can address the weaknesses of both structure paradigm and interaction paradigm theories. Social mechanism does so by integrating core elements of both of those paradigms – an acknowledgement that social reality is stratified, and an insistence that conversation represents an autonomous ontological level, mediating between individuals and emergent social phenomena.

Conclusion

The social scientists best known for emphasizing the importance of lower-level causal explanation are those who have advocated a close focus on specific cases – anthropologists explaining a specific cultural practice, historians explaining a specific historical event. It remains a risk that a mechanistic approach could provide wonderful explanations of specific token instances, but degenerate into the study of specific cases, thereby losing all generality and thus being of limited scientific value (see Mayntz 2004). After all, the social sciences that have developed the most elaborated mechanistic accounts of token social phenomena, history and anthropology, are both academic disciplines that resist lawful generalizations.

Even if lower-level causal explanation is limited in generalizability, sociologists could still develop more general scientific explanations in terms of systems and mechanisms, but the description of the system and mechanism would have to incorporate the terms and properties of macro-sociology in addition to individual properties and relations. Although "mechanism" is commonly associated with methodological individualism – because its advocates assume that a social mechanism must be described in terms of an individual's intentional states and relations (e.g. Elster 1998) – there is no reason why science cannot include systems and mechanisms at higher levels of analysis (Wight 2004). After all, individual properties such as intentional states are themselves realized in the lower-level substrate of neurons and their synaptic connections. Non-reductive individualism takes the argument that individual properties should be allowed in scientific explanation, and develops a similar argument to defend social properties and social explanation (Sawyer 2002b).

I have argued that conversational groups meet many of the criteria for emergent complex systems. However, a sustained argument that improvisational performances are complex systems requires a more extended presentation, and this is not the place for such an argument (although see Sawyer 2003b). Ultimately, the success or failure of micro-explanation in terms of individuals must be determined through empirical research. The social mechanism approach is valuable exactly because it provides a framework that allows us to carry out this exploration.

REFERENCES

Alexander, Jeffrey C., Bernhard Giesen, Richard Münch and Neil J. Smelser. 1987. *The Micro–Macro Link*. Berkeley: University of California Press.

Archer, Margaret S. 1995. *Realist Social Theory: the Morphogenetic Approach*. New York: Cambridge University Press.

Aronson, Jerrold L., Rom Harré and Eileen Cornell Way. 1995. *Realism Rescued: How Scientific Progress is Possible*. Chicago: Open Court.

Bechtel, William. 2001. "The compatibility of complex systems and reduction: a case analysis of memory research," *Minds and Machines* 11: 483–502.

Bechtel, William and Robert C. Richardson. 1993. *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Princeton University Press.

Beed, Clive and Cara Beed. 2000. "Is the case for social science laws strengthening?" *Journal for the Theory of Social Behaviour* 30: 131–53.

Bhaskar, Roy. 1975/1997. *A Realist Theory of Science*. New York: Verso Classics.

Blumer, Herbert. 1962. "Society as symbolic interaction," in A.M. Rose (ed.), *Human Behavior and Social Processes: an Interactionist Approach*. Boston: Houghton Mifflin, 179–92.

Bourdieu, Pierre. 1972/1977. *Outline of a Theory of Practice*. New York: Press Syndicate of the University of Cambridge. Originally published as *Esquisse d'une théorie de la pratique*. Genève: Droz, 1972.

Bunge, Mario. 2004. "How does it work? The search for explanatory mechanisms," *Philosophy of the Social Sciences* 34: 182–210.

Collins, Randall. 1981. "On the microfoundations of macrosociology," *American Journal of Sociology* 86: 984–1014.

Craver, Carl. 2001. "Role functions, mechanisms and hierarchy," *Philosophy of Science* 68: 31–55.

2002. "Interlevel experiments and multilevel mechanisms in the neuroscience of memory," *Philosophy of Science* 69: S83–S97.

Duranti, Alessandro and Charles Goodwin. 1992. *Rethinking Context: Language as an Interactive Phenomenon*. New York: Cambridge University Press.

Elster, Jon. 1998. "A plea for mechanism," in Peter Hedström and Richard Swedberg (eds.), *Social Mechanisms: an Analytical Approach to Social Theory*. New York: Cambridge University Press, 45–73.

1989. *Nuts and Bolts for the Social Sciences*. New York: Cambridge University Press.

Epstein, J.M. 2006. *Generative Social Science: Studies in Agent-based Computational Modeling*. Princeton University Press.

Fairclough, Norman. 1995. *Critical Discourse Analysis: the Critical Study of Language*. New York: Longman.

Fodor, Jerry A. 1974. "Special sciences (or: the disunity of science as a working hypothesis)," *Synthese* 28: 97–115.

1989. "Making mind matter more," *Philosophical Topics* 17: 59–79.

Foucault, Michel. 1969/1972. *The Archeology of Knowledge and the Discourse on Language*. New York: Pantheon Books. Originally published as *L'Archéologie du Savoir*. Paris: Editions Gallimard, 1969.

Giddens, Anthony. 1984. *The Constitution of Society: Outline of the Theory of Structuration*. Berkeley: University of California Press.

Glennan, Stuart S. 1996. "Mechanisms and the nature of causation," *Erkenntnis* 44: 49–71.

Habermas, J. 1987. *Theory of Communicative Action*. Boston: Beacon Press.

Hedström, P. 2005. *Dissecting the Social: On the Principles of Analytic Sociology*. Cambridge University Press.

Hedström, Peter and Richard Swedberg (eds.) 1998. *Social Mechanisms: an Analytical Approach to Social Theory*. New York: Cambridge University Press.

Humphreys, Paul. 1997. "How properties emerge," *Philosophy of Science* 64: 1–17.

Jackson, Frank and Philip Pettit. 1992. "Structural explanation in social theory," in Kathleen Lennon and David Charles (eds.), *Reduction, Explanation, and Realism*. Oxford: Clarendon Press, 97–131.

Kincaid, Harold. 1997. *Individualism and the Unity of Science*. New York: Rowman & Littlefield.

Knorr-Cetina, Karin D. and Aaron V. Cicourel. 1981. *Advances in Social Theory and Methodology: Toward an Integration of Micro- and Macro-Sociologies*. Boston: Routledge & Kegan Paul.

Layder, Derek. 1990. *The Realist Image in Social Science*. New York: St. Martin's Press.

Little, Daniel. 1998. *Microfoundations, Method and Causation: On the Philosophy of the Social Sciences*. New Brunswick: Transaction Publishers.

Macdonald, Graham and Philip Pettit. 1981. *Semantics and Social Science*. London: Routledge.

Machamer, Peter, Lindley Darden and Carl F. Craver. 2000. "Thinking about mechanisms," *Philosophy of Science* 67: 1–25.

Markovsky, Barry. 1997. "Building and testing multilevel theories," in Jacek Szmatka, John Skvoretz and Joseph Berger (eds.), *Status, Network, and Structure: Theory Development in Group Processes*. Palo Alto: Stanford University Press, 13–28.

Mayntz, Renate. 2004. "Mechanisms in the analysis of social macro-phenomena," *Philosophy of the Social Sciences* 34: 237–59.

Pêcheux, Michel. 1982. *Language, Semantics, and Ideology*. New York: St. Martin's Press. Originally published as *Les vérités de la Palice: linguistique, s'émantique, philosophie*. Paris: François Maspero, 1975.

Sawyer, R. Keith. 2002a. "A discourse on discourse: an archeological history of an intellectual concept," *Cultural Studies* 16: 433–56.

2002b. "Nonreductive individualism, Part I: supervenience and wild disjunction," *Philosophy of the Social Sciences* 32(4): 537–59.

2003a. *Group Creativity: Music, Theater, Collaboration*. Mahwah: Erlbaum.

2003b. *Improvised Dialogues: Emergence and Creativity in Conversation*. Westport: Greenwood.

2003c. "Nonreductive individualism, Part II: social causation," *Philosophy of the Social Sciences* 33(2): 203–24.

2005. *Social Emergence: Societies as Complex Systems*. New York: Cambridge University Press.

Schegloff, Emanuel A. 1992. "In another context," in A. Duranti and C. Goodwin (eds.), *Rethinking Context: Language as an Interactive Phenomenon*. New York: Cambridge, 191–227.

Stinchcombe, Arthur L. 1991. "The conditions of fruitfulness of theorizing about mechanisms in social science," *Philosophy of the Social Sciences* 21: 367–88.

Tannen, Deborah. 1993. *Framing in Discourse*. New York: Oxford University Press.

Turner, Jonathan H. 1993. *Classical Sociological Theory: A Positivist Perspective*. Chicago: Nelson-Hall.

Wight, Colin. 2004. "Theorizing the mechanisms of conceptual and semiotic space," *Philosophy of the Social Sciences* 34: 283–99.

Yablo, Stephen. 1992. "Mental causation," *The Philosophical Review* 101: 245–79.

Part II

Mechanisms and causality

5 Generative process model building

Thomas J. Fararo

Introduction

Theoretical sociology deals with theoretical problems and methods, formulating principles and models that incorporate generative mechanisms in applications to abstract or concrete cases within a specified scope (Fararo 1989). This approach is "generalizing" rather than "historical" or particularizing in its orientation (Berger *et al.* 1972). Part of what this means is that the importance of an empirical study is tied to the theoretical problem in the context of explanatory theory development. This viewpoint has governed the growth of a variety of theoretical research programs in sociology (Berger and Zelditch 2002). Theory-driven studies of cooperation, such as those stimulated by the pioneering work of Axelrod (1984), are in this spirit.

Given the generalizing orientation, we can distinguish between present-day embodiments of it and certain mid-twentieth-century efforts that are often grouped under the rubric "general theory." I am referring to not only Parsons but also, by contrast, Homans. Their earlier efforts differ from present-day contributions in that they had the aim of *general theoretical synthesis*. They aspired to create a solid theoretical foundation for not only sociology but the other social sciences as well. Parsons initiated his synthesis efforts with a focus on the problem of social order. Less obviously, this was true of Homans as well. The order problem in his theoretical works is transmuted into a theory of spontaneous order – how structures of stable social relations emerge in social interaction, as in the treatment of what he called "the internal system" (Homans 1950). His synthesizing objective is with respect to empirical findings grounded both in empirical field research and in experimental studies (Homans 1958, 1974 [1961]). The explanatory arguments, although couched in terms of behavioral psychology, incorporate theoretical ideas from the wider social science tradition.

For Parsons, by contrast, the general problem is explicitly discussed in terms of its classical Hobbesian formulation in the context of defining a

sociological agenda grounded in a critical analysis of the ideas of major social theorists in the sociological tradition (Parsons 1937). Over decades of conceptual construction and reconstruction, Parsons remained committed to the overriding synthesizing objective as he drew upon contributions from psychology, anthropology and other fields, including cybernetics with its focus on communication and control.

Although these two general theorists endeavored to implement an analytical approach in sociology their theoretical models were deficient in what I call *generativity*. As concisely summarized by Manzo (2007: 44), whose paper relates this notion to many recent developments in sociology, this means that "attention is focused on the emergence, engendering or genesis of what is observed." Starting from the idea of a generating process and the associated strategy of constructing theoretical models that combine mechanisms (Fararo 1969, 1972), my ideas evolved into what I called "generative structuralism" (Fararo 1989; Fararo and Butts 1999), the combination of generativity with a focus on social structure. Examples of both the earlier and later ideas just mentioned will appear in this chapter. But first, an overview.

In the first section of the chapter, I elaborate on my discussion of Homans and Parsons in regard to their formulation of two key generalizing orientations in mid-twentieth-century sociology with a focus on their analytical approaches. I then make a transition to discuss the generative mechanism style of theoretical model building. The second section discusses mathematical models that incorporate some notion of a generating process or mechanism, either in a stochastic or deterministic mode, with special reference to models with unobservable or "hidden" state spaces. In the example to be discussed in some detail that deals with the problem of the emergence of hierarchy, the additional feature is that the state of the system is a dynamic social relational configuration. The third section discusses generativity in the context of certain developments in cognitive science that drew upon the mid-twentieth-century cybernetic zeitgeist. But the analytical focus is sociological as sketched in two types of models. The first type represents institutional structures as systems of distributed generative rules that are implicit in the actions and interactions of the actors. The second type takes up the problem of how class-dependent subjective images of a stratified system are generated as unintended by-products of social interactions embedded in that system. The fourth section concludes the chapter with a discussion of generativity in computational sociology. An example combines certain elements of models discussed in earlier sections by dealing with both the emergence of hierarchy in a social network and the emergence of subjective images of that same network in a simultaneous

generative process. Given the context of an article-length treatment, of all the examples discussed in what follows only one will include formal details.

Analytical theorizing in post-classical perspective

In the decades in the mid-twentieth century from the 1930s to the 1970s, the search for a general theoretical framework for the analysis of social phenomena was at its high point. Sociological theory was entering its post-classical era. In the Harvard environment of the mid-1930s, certain intellectual influences helped to foster this ambitious aspiration of Talcott Parsons and George Homans. The novice social scientists were particularly taken with Alfred North Whitehead's *Science and the Modern World*, which had appeared in 1925. As interpreted by them, Whitehead was stressing the importance of a *conceptual scheme* for a science. For instance, in his first influential book Homans (1950) would start by explicitly stating that certain specified terms comprised a conceptual scheme necessary but not sufficient for a theory of the human group as a social system. At about the same time, Parsons (1951) also set out what he referred to as a conceptual scheme for a social systems theory. Parsons was particularly attuned to another idea he drew from Whitehead: the importance of avoiding a potential problem in employing a conceptual scheme, namely coming to regard the explanation of the behavior of a concrete system as exhausted by the analytical categories of that scheme. This problem was dubbed "the fallacy of misplaced concreteness" by Whitehead in his 1925 book and cited repeatedly in later years by Parsons.

Thus, in their then common view, for Homans and Parsons a conceptual scheme is the starting point for an *analytical theory*, a type of theorizing they learned about in attending the famous Pareto seminar in the 1930s, where they were no doubt exposed to Pareto's outline of the logic of the analytical method in scientific theorizing as well as to his aspirations for a general sociology using that method. (See Fararo (1989: Sect. 2.6) for a detailed discussion of Pareto's methodological ideas.) Each embarked upon a program for the building of a general analytical theory as a *system of analytical laws* wherein each such law states a relationship among *analytical elements* specified in the associated conceptual scheme.

In *The Human Group*, Homans approximates this procedure. His conceptual scheme consists of three categories that help him specify associated features that function as analytical elements, including (but not limited to) frequency of *interaction*, intensity of positive *sentiment*

and similarity of *activity*. He then proceeds to frame an analytical theory consisting of statements he calls analytical hypotheses – holding back on calling them laws until further evidence supports them as having that character – that are necessarily interrelated by utilizing the same elements in different relations. In a chapter devoted to the analysis of social control in groups, it is made clear that the ideal form of an analytical theory of a dynamic social system would be a system of differential equations. This idea has the most credibility in physical theory. A mathematical physicist and philosopher of science Henry Margenau (1950) came to a similar conclusion after an impressive conceptual analysis of the formal representations employed in a variety of fields of physics. (He notes that almost all general physical theories involve systems of *partial* differential equations because change of physical quantities is analyzed with respect to the element of space as well as time.) However, for various reasons it is unlikely that this sort of representation will ever be common or even appropriate for sociology.

In his later work, Homans (1974 [1961]) shifted to the idea of an analytical theory as a deductive system (as he found it set out by Braithwaite (1953)) and adopted a conceptual scheme grounded in behavioral psychology. Roughly speaking, the mechanisms specified by the hypotheses of the earlier theory are now themselves to be accounted for by more primitive mechanisms. For instance, the hypothesis that interaction and positive sentiment are mutually dependent – an increase in either produces an increase in the other, with activity held constant – calls out for explication and qualifications. By introducing what amounts to a reinforcement mechanism (called the "success proposition") and treating interaction as "exchange" in which the acts of each actor reward and/or punish the other, Homans simultaneously explicates, explains and qualifies the original statement so that the interaction-liking relationship in the process of the build-up (or "build-down") of the social system of the group is not false but neither is it a law. It is a *mechanism* whose scope of effect becomes a matter of contingent events – as framed in terms of the behavioral elements of the more primitive analytical scheme. From observations of external behaviors an analyst might venture to code an over-time sequence of interactions in terms of the categories of reward and punishment but this is not necessarily a readily accomplished task. In any case, since few sociologists were willing to take this path of research, the theory became embalmed in textbooks as "Homans' exchange theory" and was treated as a kind of intellectual prelude to the recent work of some experimental sociologists who have developed "exchange network theories." These theories are not closely

related (and even in some cases antagonistic) to Homans's own analytical theorizing.

Consider now how Parsons implemented the dual ideas of conceptual scheme and analytical theory in the context of the ambition he shared with Homans to create a *general* theoretical foundation for scientific sociology. While Homans was not altogether neglectful of classical sociology – citing not only Pareto but also Simmel, for instance – Parsons (1937) forged his approach in terms of what he saw as the *convergence* of certain classical writers on a common conceptual scheme, the key feature of which is the inclusion of the "end element," that is, the purposive character of human action, and the accompanying "normative element." Taken together, purposes and norms serve to create a kind of early version of cybernetic hierarchy in which upper levels involve the most general ends (called values) that are specified in chains of contextualization that terminate in behavioral goals in situations. Parsons links this hierarchy concept to the problem of social order that he frames in a discussion of Hobbes in which he argues that a necessary, but not sufficient, condition for social order is a *common value system*.

The prominence of values as highest-level controls in the hierarchy implies that to flesh out this conceptual scheme requires some mode of analysis of common value patterns found in concrete social systems. In the 1951 treatise *The Social System* and in other works appearing around that time, Parsons specifies a set of "pattern variables" for the analysis of values, especially institutionalized value patterns illustrated by the factual statement that American society institutionalizes the achievement side of the achievement-ascription pattern variable.

Whatever the merits of the pattern variable conceptual scheme and the approach to the problem of order adopted in relation to it, Parsons did not pursue the transition to an analytical theory (of social order) as a system of analytical laws as he had suggested was on his agenda in *The Structure of Social Action*. His convergence thesis was with respect to structural parts and relations, not with respect to such an analytical theory, but the latter was to be the desired payoff of adopting the "voluntaristic" action frame of reference with its purposive and normative elements. Instead, in the interim between the 1937 work that set out the classical foundations for the general action conceptual scheme as a kind of cybernetic hierarchy and the 1951 social system analysis, Parsons must have grasped the many difficulties involved in creating a wide-scope analytical social theory interpreted in his earlier terms. It seems that, like Homans, he sought some alternative mode of theory construction. Whereas Homans considered his shift to a deductive system grounded in propositions from behavioral psychology as a scientific

advance, Parsons describes his shift to *structural-functional analysis* as a "second-best" form of theorizing (Parsons 1951: 20).

The key conceptual step in initiating this approach, he argues, is to take the institutional pattern of the social system as given, so the explanatory task becomes: how is such a pattern maintained? In terms of the classical mathematical analysis of systems, employed by Parsons in an analogical mode, the focus is on the problem of stability of equilibrium, treating social change as instability leading to a new equilibrium. Parsons treats this problem by specifying two classes of mechanisms relating to socialization and social control. The first class pertains to the internalization of institutionalized values that are the highest levels of cybernetic control and the second class arises because socialization is not sufficient to prevent deviations from institutionalized patterns.

Thus, it would be quite incorrect to state that Parsons did not have a mechanism-based theory. But it would be correct to state, as I do, that his theoretical explanations do not exhibit generativity. Admittedly, my criterion is closely aligned to a formal process representation that Parsons did not formulate. Instead, as his structural-functional form of analysis became the well-known "four-function scheme," there appeared the problem of specifying the relationship between it and the pattern variable scheme, a problem addressed in a series of publications that culminated in a paper (Parsons 1960) that most readers probably found completely baffling in both its textual discussion and its accompanying quasi-formalistic diagrams.

Contextualization and embeddedness are built-in features of Parsons' mode of structural-functional analysis because any entity – actor, situation or whatever – must be "located" in the scheme and then itself recursively analyzed in terms of the four functions and their interconnections. Parsons' later work is replete with discussions of mechanisms in these contexts. In this sense, Parsons – like Homans – is an analytical process theorist. But his formalism is not adequate to the task and so his theoretical models lack generativity. For more on this point and a more extended explication, analysis and critique of the theoretical systems of Homans and Parsons in the wider context of the heritage of general theory in sociology see Fararo (2001).

In this section, I have provided a sympathetic treatment of how two mid-twentieth-century sociologists, Homans and Parsons, attempted to create general theoretical systems that were synthesizing, explanatory and analytical in methodological orientation. These aims proved very difficult to realize in combination. Homans came closer in terms of his adoption of a deductive approach but not in his choice of behavioral psychology as sufficient for the task. In any case, behavioral psychology

was already being transcended by developments in cognitive science (see below, pp. 111–14).

Very importantly, the lack of effective formal methods for the pursuit of their theoretical goals was particularly clear in their treatment of process. Namely, their theoretical models – the particular explanatory arguments they gave in application of their general theoretical frameworks – lacked generativity. What this means will be illustrated in the remaining sections of this chapter but an initial orientation to the approach may be appropriate at this point. Every generative process model has a causal aspect in that it shows how varying outcomes arise out of varying conditions under which the process occurs. But the generative process approach, as we will see in what follows, differs from "causal model building" in that the process that accounts for the linkage between conditions and outcomes is formally specified. There is a mathematical or computational series of process steps or transitions of the state of a system that explicitly show how starting from certain conditions the generative rules produce the outcome and thereby *explain* the causal linkage. Thus *temporality* is essential to the generative process approach and this differs from only stating a causal connection in words or in an arrow diagram. (For a further discussion of the concept of causality in relation to generative process model building, see Fararo (1989: Sect. 2.10)).

Generativity: mathematical models

As is widely understood, a mathematical model of a process can take a stochastic or a deterministic form. In this section, I discuss and illustrate these two forms of process models and then refer to one of my own research programs to provide an extended example of generativity in the stochastic process context.

Deterministic process models

A good example of a deterministic process model is Simon's (1952) formalization of the theory of social systems set out by Homans (1950) in terms of a system of differential equations. Generativity is a keynote feature of this classical form of applied mathematics. In the physics of the motion of a satellite, for instance, an orbit is not just there as a static form but is *generated* by the concatenation of forces acting on the moving object at each instant. These forces are represented in a dynamic model satisfying a Newtonian template for such model building. Recall that Homans' theory consists of a set of hypotheses, each linking a subset

of the same small set of analytical elements. If the phenomenon to be explained is the emergent social structure of the group or the change of its given structure, each hypothesis can be interpreted as specifying one mechanism entering into a theorized combination that produces such outcomes under certain conditions.

The sheer number of mechanisms to combine would introduce formidable analytical complexity so it is not surprising that Simon's model constitutes a highly simplified edition of the theory. Using the strategy of initial simplification, Simon starts from a linear model, but then goes on to formulate and study the consequences of a nonlinear model. Basic questions about such a system of social processes comprising a human group can be addressed by methods for analyzing such models. For instance, does the dynamic system have any equilibrium states and, if so, how many and with what properties, such as stability? These are the same questions that both Homans and Parsons had been addressing by a kind of analogical mode of reasoning, as I pointed out earlier.

It is this ability to provide us with mathematical methods for the analysis of both stability and change as framed within a process approach that is the attraction of such dynamic deterministic models. They are generative in the sense that often one can solve the system for the implied trajectories – the over-time changes of state in the state space (analogous to the orbits mentioned earlier). Nonlinear systems for which explicit solutions cannot be derived can be studied by methods of approximation. Finally, very interesting further analytical studies of nonlinear dynamics can be made that focus on such phenomena as abrupt changes of state. Thus, in principle, such a "dynamical system" formulation can be the basis for the analytical study of key problems in sociological theory dealing with the emergence, stability and change of social structures (Fararo 1989: ch. 2). The complexity problem, however, limits the effectiveness of analytical methods in the derivation of system behavior and this leads to computational methods, as discussed and illustrated in the last section of this chapter.

Stochastic process models

In formulating a stochastic process model, we think of a system that can occupy a number of states such that there are transitions between states that repeatedly occur through some probabilistic mechanism. If the mechanism does not require reference to the past history of the process it is called Markov. A fundamental fact about processes in sociology is that human psychological and social processes that are represented in terms of observable behaviors or conditions are usually *not* Markov.

However, as illustrated in a model of the famous Asch situation constructed by Cohen (1963), a key strategy – not always workable – is to postulate a space of unobservable or hidden states such that the transitions among such states have the Markov property as part of their systematic meaning. Then the logical structure of the model takes the form of a set of assumptions about this Markov chain of unobservable states and a further assumption connecting such states to observable behaviors.

My assumption has been that in conjunction with this strategy Markov processes could play a key role in sociological model building, as illustrated in an early work (Fararo 1973) and as further instantiated in the construction of models that combine this feature with a focus on the dynamics of a social relational configuration, as I now discuss by way of illustration of the stochastic process mode of generativity in sociology.

Example: E-state structuralism

What we have called the *theoretical method* of E-state structuralism (Fararo and Skvoretz 1984) was proposed initially as a way of integrating core elements of two research paradigms that to that point in time had evolved separately, namely social network analysis and expectation states theory (Berger and Zelditch 1985).

The social network paradigm is characterized by its focus on patterns of social relations constituting a network of ties between nodes representing people, organizations or even animals in some studies. For example, Chase (1980) undertook the observation of groups of animals with a focus on the problem of the emergence of hierarchy, an important structural form that may arise in the dynamics of social networks.

The nature of expectation states theory, the second of the two paradigms underlying E-state structuralism, is less well known. Its originator is Joseph Berger in collaboration with Morris Zelditch, Jr. and Bernard P. Cohen. These sociologists did their doctoral studies at Harvard in the 1950s, the decade of Parsons' maximal influence. It was a decade in which the focus on interpersonal processes was prominent in the work of other faculty members, including Homans and Bales. Expectation states theory, as it was initiated in this environment, dealt with social influence among members of groups studied under controlled conditions.

The theorists argued that sociological theories were to be abstract and general, in agreement with Parsons. But by and large expectation states theory is based upon a negative evaluation of the corpus of

Parsons' work. Berger and his colleagues, together with many generations of students who have contributed to the development of the theory, argue that sociological theories have to be framed in unambiguous terms. Elements such as definitions, assumptions and derivations must be clearly distinguished. Theoretical arguments must be associated with stated scope conditions so that derivations can be tested under appropriate empirical conditions. Successful derivations support specific theoretical arguments and lead to efforts of integration of specific theories in the theoretical research program constituting expectation states theory.

This program continues to produce and develop a system of theories employing a common set of general concepts and principles pertaining to the acquisition and operation of social expectation states. The paradigm has employed both formal theorizing and experimental methods that are theory-driven. An important methodological idea is that the experimental situation is only an instantiation of a generic type of situation within the stipulated scope of a particular specific theory. This includes the relevant social status characteristics that may be activated in the situation in the sense of becoming salient for the actors. So, for instance, the theory uses the term "diffuse status" and specifies mechanisms relating to its functioning (or not) in a group process. In particular instances, whether experimental or otherwise, this theoretical concept may be instantiated by gender, class, race, educational level or any other dimension that satisfies the abstract definition of the concept. Similarly, the mechanisms associated with the operation of one or more diffuse statuses in a situation are framed abstractly. Each specific theory is set out along with its stipulated abstractly stated scope, such as collective problem-solving situations in which the actors are initially unacquainted. Experimental realizations are constructed to be within the scope while providing the opportunity to test the specific abstract theory. If a theoretical prediction fails in that context, the theory needs revision; if it succeeds, this is evidence in its favor, although not conclusive (in the spirit of science). In the nature of most experimentation, repetitions of the social process of interest are part of the design and from the standpoint of a mathematical model of the process this gives rise to the associated idea of a stochastic process model that can be tested in terms of the aggregate experimental data.

E-state structuralism (Fararo and Skvoretz 1984) is a method for setting up and studying Markov process models with unobservable "E-states" that applies to both animal studies of the type that Chase and others undertake and human studies of the type that expectation state theorists undertake. In either case, we must distinguish between

social relational expectations (not observable) and social behaviors (observable). The E-states are interpretable as dispositions for behavior of each member toward specific others or toward generalized others (e.g. all those typified in the same way) and are stipulated to emerge and change in social situations.

In this type of model *the system state is a matrix of relations among all pairs*, in which each relation is defined in terms of a conjunction of E-states. This shift to a structural state description, while preserving and generalizing the logic of the expectation states construct, is what justifies calling this procedure E-state structuralism. Prior to this procedure, expectation states theorists had no method by which they could generate the evolving structure of expectations in the situations they studied. By linking the ideas of this theoretical approach to social network thinking, E-state structuralist models incorporate mechanisms that together generate a trajectory in a space of possible E-state matrices along with manifestations of states in the observable social behaviors of the actors. In particular, a key mechanism involves third parties who witness a given pair interaction and whose E-states may emerge or change on the basis of this bystander role in the immediate interaction situation. The initial E-state model pertained to the problem of showing how emergent hierarchy is generated in small groups of initially unacquainted animals and for which we proposed an E-state interpretation of the bystander role that Ivan Chase had proposed would explain emergence in such cases.

The procedure for constructing an E-state model has a constructive phase in which assumptions or axioms specify the process and a deductive (and/or, often, a simulation) phase. As indicated above, in the example under discussion, the constructive phase proposed an E-state interpretation of the bystander mechanism to explain the emergence of hierarchy. The relevant observable behaviors are aggressive actions called "attacks." The assumptions describe the probability of the formation of E-states, given an attack event, both for the immediate parties to the attack and any bystanders. When animal α attacks animal β, both may change E-state: α to a dominant state, β to a deferential state, and if bystander γ observes the attack, a further mirroring may create E-states in which γ is in a deferential state toward α and in a dominant state toward β. Another assumption sets out a constraint on attack events, given the current structure of E-states. For instance, if animal α has a deferential E-state toward animal β, then α will never attack β, although β may attack α. The formal representation is sketched briefly as follows.

We begin with a definitional matter. We interpret a *dominance relation* as a conjunction of two complementary relational states: first, animal α

has a dominant E-state toward β (and we write $\alpha E_H \beta$) and, second, animal β has a deferential E-state toward α (and we write $\beta E_L \alpha$). Formally, for any α, β:

$\alpha D \beta$ if and only if $\alpha E_H \beta$ and $\beta E_L \alpha$

This relation is characterized by the condition: whenever the "high" or dominance relational state forms for one animal toward another, the corresponding "low" or deferential state forms in the other animal as a consequence of a particular encounter. This is not inevitable and, in fact, is only one possibility among others treated in subsequent work, but it simplifies the initial model. The axioms defining this simplest model of dominance structure formation are:

Axiom 1. (Initial Condition). At t = 0, every pair is in state not-D.
Axiom 2. (E-state Formation). At any t, if a pair is in state not-D and if one member α attacks another β, then $\alpha D \beta$ forms with probability π.
Axiom 3. (E-state Stability). Once D forms, it is retained: for any α, β and time t, if $\alpha D \beta$ at t, then $\alpha D \beta$ at t+1, no matter what attack event occurs at t.
Axiom 4. (Deference). At any t, if $\alpha D \beta$, then β does not attack α at t.
Axiom 5. (Bystander). At any t, given an attack event in which some α attacks some β, then their relationships to any third animal γ, called a bystander, may change as follows:
if α(not-D)γ at t, then $\alpha D \gamma$ at t+1 with probability θ
if γ(not-D)β at t, then $\gamma D \beta$ at t+1 with probability θ and these events are independent and also independent of the event in Axiom 2.
Axiom 6. (Attack Events). At any t, given the constraint of Axiom 4, all potential attacks have the same probability of occurrence.

This postulated process is a Markov generating process. Note that the observable behaviors, the attacks in this model, are probabilistically dependent on the system state but the system state changes over time, depending on whether or not (in terms of the probability parameters) E-states form in the aftermath of particular attacks. But since the process is shown to be an absorbing Markov chain, there are E-states that so constrain which attack actions will take place at all and those that so validate the existing E-states that no further change occurs. These states are all either cycles or hierarchies. When a derived formula for the probability of a hierarchy is analyzed, it is found that for most parametric conditions, the probability is nearly unity that the process is absorbed into a hierarchy, which is essentially the empirical regularity to be explained when this regularity is interpreted as an equilibrium of a generating process.

An empirical test of the model requires estimation of the parameters using appropriate data. By analogy with the mathematical notion of a

function of several variables, the model is a function of two parameters. When these are estimated in a re-analysis of Chase's data (Skvoretz and Fararo 1988), we find that the model does less well in the generation of the time path to equilibrium. This suggests the need to revise certain elements of it. Chase and Lindquist (2009) assess various models including the E-state model described here in the context of providing the most relevant current statement on the emergence of hierarchy problem in animal groups with special reference to Chase's theoretical ideas and empirical studies.

A subsequent application of the E-state structuralist method deals with the explanation of the emergence of status orders in human task groups (Skvoretz and Fararo 1996). The observable acts are more varied but largely verbal as dominance takes the more subtle human form. In other respects, key ideas and techniques of the simple model are carried over to this problem, including the dynamic approach, the matrix state space and the bystander mechanism. However, these features are embedded in the conceptually rich theoretical framework of expectation states theory – as contrasted with the stripped-down features of the simpler model – to produce an integrated theoretical model that generates a variety of types of outcomes as a function of various parameters. An empirical test of the model is reported by Skvoretz *et al.* (1999).

Generativity: cybernetic models

Just as Homans worked out his theories in an intellectual environment dominated by behaviorism, later theoretical work has drawn upon cognitive psychology and, more broadly, cognitive science. Formal developments in this field have had an important impact on sociology (Bainbridge *et al.* 1994). Most relevant here are several works that reflect the post-Second World War cybernetic zeitgeist.

Miller *et al.* (1960) applied the concept of negative feedback to create a theory that linked the actor's dynamic situational knowledge to that actor's behavior under the assumption that the actor engages in purposive action. The generic dynamic unit, called a *plan*, generates an act as an operation to reduce the difference between a goal and information about a situation. In a model consisting of levels of plans, all except the most general goals are set via control at a higher level. (Here we see a conceptual link to Parsons' action theory.) This idea of a cybernetic hierarchy of negative feedback units was developed further by Powers (1973) both in the "upward" direction in which goals are ultimate ends and in the "downward" direction in which goals relate to physical movements.

Such behavioral control system thinking has been drawn upon in a number of contemporary research programs in sociology (McClelland and Fararo 2006). For example, McPhail *et al.* (2006: Fig. 3.2) explain crowd behavior with a theoretical formulation that includes specification of a model consisting of a three-level cybernetic hierarchy in which there is an uppermost *meta-symbolic level* of values and beliefs, an intermediate *symbolic level* of plans and programs, and a lowest *pre-symbolic level* of physical movements.

The symbolic level is treated in enormous detail by Newell and Simon (1972). These authors make the concept of *information processing system* central to their analysis of human problem solving, especially complex problems requiring considerable thought. There are symbolic expressions that constitute declarative knowledge about something, whether simple or complex, concrete or abstract. And there are symbol expressions that constitute if-then rules and systems of them. These comprise *programs* that, via control over the lowest level of the cybernetic hierarchy, yield physical acts – behaviors with meanings arising from the symbolic and meta-symbolic levels.

What follows are sketches of two collaborative research programs in which the ideas and methods in each case grew out of these influential examples of cybernetic control theories in the context of treating sociological problems.

Example: generating institutionalized social action

The method of Newell and Simon involves writing computer programs that model the production of actions taken by people working on cognitive problems, such as playing chess or solving nontrivial puzzles. In their experimental settings, problem solvers talk aloud as they think about steps they might and do take in solving such a problem, leading the analysts to argue that each solver works within an implicit "problem space" that enables and constrains some possible knowledge states and their linkage to actions. Based upon the data, for each problem and problem solver, the model includes both a postulated problem space and a hypothesized program comprised of a set of if-then rules. Each rule takes the form of specifying an action in the context of a goal and a state of knowledge in a particular context. Here we see the dynamic link between declarative situation knowledge and the instructional or action-oriented type of knowledge mentioned earlier. Newell and Simon point out that such a program is functionally equivalent to a system of differential or difference equations in that it is a model that accounts for a succession of changes in the behavior of a system. In the

language of this chapter, it is a generative model. Newell and Simon show that in most cases they could generate the observable sequence of context-dependent actions with a high level of accuracy.

But of what relevance is all this for sociological theory?

Berger and Luckmann (1966) drew upon classical sociological thinking to formulate a linkage of a generalized institution concept to the problem of social reality in the writings of the social phenomenologist Alfred Schutz. A similar generality of the concept had been more or less implicit in the writings of some other social scientists, such as Nadel (1951) who explicitly defined institutions in terms of if-then rules. The new element in the Berger–Luckmann treatment of the concept was the strong cognitive emphasis in a shift from the sociology of intellectual knowledge to the sociology of everyday knowledge involving "typification" schemes (types of actors, acts and situations) that constitute everyday behavior as institutional. Once an institution emerges, it counts as part of social reality in the sense of "what everybody knows" is true of their social environment whether they like it or not.

With the publication of the Newell and Simon book in 1972, a group of us adapted their mode of representation to sketch a generative model of a specific institution and to suggest it as a means toward the goal of creating a formal theory of institutions (Axten and Fararo 1977). The generativity consisted of deriving specific interconnected sequences of actions of multiple actors from the postulated structure of rules comprising the model. The important new aspect of such model-building, relative to that of Newell and Simon, is the strong interdependence of the situations and actions of the various actors comprising a *social* system. (See also Heise (1989) for a related approach.)

The core of the model pertains to the symbolic level of cybernetic hierarchy and employed the Newell–Simon formalism to represent typification schemes and related properties of institutions. The role of social symbols is fundamental. A common space of symbols is presupposed for an institutional domain and the behavior of each actor is under the cybernetic control of what we called a *rolegram*. Each rolegram relates to other rolegrams to form a template or grammar for process. An activated template along with possible concrete choices made by the actors generates *the normal forms of interaction* in an institutional setting. These normal forms are like the syntactically correct sentences of a generated language while the system of rolegrams is like a grammar for that language. The number of rolegrams is much smaller than the number of actors because multiple actors perform the same function in the institution and have the same role name in the associated social symbolism. As in generative linguistic models this type of institution model does

not explain the choices made by the actors. Rather it specifies the social context for *such* a choice, i.e. one that will count, institutionally, as a valid instantiation of the type of action called for in the particular state of the situation as defined in institutional terms.

Fararo and Skvoretz (1984) provide a lengthy exposition and illustration of the theoretical approach just sketched and in a subsequent paper a more critical analysis of the limitations of the approach (Fararo and Skvoretz 1986). For instance, there is no effective formal treatment of institution emergence, a problem addressed by Skvoretz and Fararo (1995) in sketching an initial model. Another possibility is to treat institution emergence in problem solving terms using an approach close to that of Newell and Simon in this respect but now with a social problem space shaped by values and beliefs at the meta-symbolic level. However it also will be important to link this effort to the "subinstitutional" level of actor reasoning. Work in cognitive science that has its heritage in the Newell–Simon approach is quite relevant (Anderson 1993), as are the new developments in computer science that relate to distributed artificial intelligence that are usually grouped under the rubric "multi-agent systems" (Woolridge 2002).

Generating images of social structure

The previous discussion focused on how a formal treatment of cognition at the symbolic level of cybernetic hierarchy is relevant to sociological ideas about social reality when these ideas are embodied in a research program that employs the method of generative process model building.

A second example will again relate to social reality as understood in cognitive sociological terms but in a different mode arising out of a different theoretical problem.

Rather than representing institutions in terms of generativity at the symbolic level of cybernetic control and incorporating the typification element stressed by theorists in the reality construction tradition, this other approach takes on the problem of explaining how and why people acquire subjective images of objective structures that are shaped by social interaction.

Consider a relevant descriptive account that was presented in an anthropological study of a southern community of the United States in the 1930s (Davis *et al.* 1941). A diagram in the book showed the various class-dependent perspectives on the class system of the community. The social structuration of the entire panoply of images is what was is interesting. In treating the general problem with this account and diagram

in mind it was noted that the various different class perspectives all had a certain formal property: each image preserved the objective ordering of classes, each reduced its complexity by lumping some classes that were distinct in the objective structure, and there were formal symmetries between various images. In a word, the diagram was interpreted as (an approximation to) a *system of interrelated homomorphisms*. When the general features apparent in such a system are stated explicitly they constitute a set of abstract empirical generalizations to be explained (Fararo 1973: ch. 12). The remarks below about the explanation of these generalizations are discussed in terms of the most recent state of the theoretical framework (Fararo and Kosaka 2003).

Applying the theoretical and methodological perspective taken in this chapter, the explanatory strategy takes the form of constructing a model that specifies a cognitive process that is activated in social interaction such that the model *generates* the system of homomorphic images as an equilibrium state. According to the usage stated earlier, then, the cognitive process is functioning in the model as a mechanism. The details of the cognitive process and their relationship to social interaction within a class structure are set out as axioms that define the model.

The class structure is represented as a multidimensional stratification system with "lexicographic" ordering: each dimension is a mode of ordering and the dimensions themselves are ordered. In a feudal system, as discussed by Veblen (2007 [1899]: 7), the two leading dimensions would pertain to warfare and religion: warrior status (or lack thereof) and priestly status (or lack thereof). In the context of the American Deep South of the 1930s in the community study cited earlier, race is the highest basis of ranking and dimensions with internal ranking such as family background and wealth would follow. By implication, these dimensions vary in importance for the cognitive process activated in social interaction. What does the latter point mean? According to our model, upon first encountering anyone in the community, in defining the situation, a person more or less tacitly seeks information about the other, including class position. We postulate that the order of the cognitive search corresponds to the social order of the dimensions and the search process is terminated when a class distinction can be made – or until all dimensions are exhausted and so defining the other as a class-equal. The process is "satisficing" in the sense of acquiring only enough information to make this implicit decision about relative position.

Each such encounter updates the image that is built up from a hypothetical initial undifferentiated state until the state of the image is such that the process reproduces that state, which is then an equilibrium image (under the given specified conditions). There are idealizations

and approximations in the construction and application of such a model, as might be expected. For instance, because it is stated abstractly, the theoretical model could just as well apply to medieval Japan as modern Japan, as my colleague Kosaka has done and reported in our monograph. But in the modern case it is only a first approximation. Numerous extensions and refinements were introduced over a period of time as the general model was developed. One example is the introduction of social mobility. This is treated as a succession of processes in which the equilibrium image acquired in the initial position is followed by social interactions in the new position that lead to a generally more complex equilibrium. In another example of extension of the theory, Kosaka proved that with one additional assumption the axioms imply a theorem that explains the high frequency of middle-class self-classification that is found repeatedly in empirical studies.

Generativity: computational sociology

In working with theoretical models expressed in mathematical form, a convenient and very important tool for the analysis of the implications of any such model is simulation, defined as "a dynamic model implemented on a computer" (Evans 1988: 19).

In Hummon and Fararo (1995a) we returned to the simple E-state model of the emergence of a dominance hierarchy in small networks and studied the model by computer simulation methods. One aim was to simulate a process in which the actors are acquiring images of the very network in which they are embedded as the structure of that network is evolving. The situation of the actor is quite different here than that assumed earlier where we envisioned an objective multidimensional space in which an actor is embedded as an initial condition. In this study, stratification is emergent rather than given.

The major result concerning images reflects this difference in the givens. The emergent dominance structure is not completely ordered but the each of the images, which as earlier vary by position, is completely ordered. This result holds for all the group sizes considered and for all values of the parameters, in particular the magnitude of the effect associated with the bystander mechanism. A further discussion of this computational model is given in Fararo and Kosaka (2003: ch. 7).

Computer simulation is also employed in a paper that defines and studies some more complex E-state processes (Fararo *et al*. 1994). In the simple model, intended to apply to a limited range of infra-human social interaction, the observable level consists of hostile encounters called "attacks" that might or might not give rise to the more enduring E-states

pertaining to dominating or being dominated (called deference). The new models retain the specific problem focus on dominance while generalizing the initial model in two directions. First, we incorporate parallelism in attack processes, that is, multiple attacks among members of the group can occur at the same time. Second, we allow for non-complementarity in the formation of dominance or deference E-states. That is, earlier we assumed that if an aggressive encounter in which α attacks β occurs, then with a certain probability α forms a dominance orientation toward β and β forms a deference orientation toward α. This assumption of complementarity was an idealization meant to keep the first model simple. But in reality β may not "reciprocate" α's dominating orientation with a deferential orientation on the basis of the aggressive encounter. For instance, a conflict relation may arise in which each is disposed to dominate the other. The extended models can generate this type of outcome as well as mutual deference, which implies by Axiom 4 that neither initiates an aggressive action toward the other. As in the simple model, the important bystander mechanism is operative. With parallelism and non-complementary, computer simulations show that other equilibrium structural forms arise including *coalitions* in addition to the cycles and hierarchies found in the initial simple model.

In these computational models, generativity is transparent in the simulations: the program rules generate encounters, attacks, E-state formation events and so forth. As these features are iterated through the program operation in simulated time, eventually we obtain a list of who-dominates-whom from which we abstract the structural form of the outcome.

These models are illustrative of generativity via computational model building guided by theoretical concerns. There are various options involved in formulating such a computational model and one can think of these as choices from a menu (adapted from Fararo and Hummon 1994):

State space: discrete, continuous
Observables: discrete, continuous
Parameter space: discrete, continuous
Time domain: discrete, continuous
Timing of events: regular, incessant, irregular
Generative mechanism(s): deterministic, stochastic
Formal statement of mechanisms: equations, transition rules

Our models employ discrete event simulation, a conjunction of a continuous time domain and an irregular timing of events (as in the arrival process of people at a bank). The models also employ what computer scientists call object-oriented programming (Hummon and Fararo 1995a).

The use of computer simulation in our work exhibits generative structuralism, cited earlier in this chapter as a theoretical orientation that aims to combine generativity with a focus on social structure. Making explicit the basis in human action, an unpacking of the generative and structuralist features of this orientation leads to the following explication:

> "Generative" refers to the social behavior of actors and, in particular, to how the concatenation (interaction or interdependence) of their actions generates collective outcomes. "Structuralism" refers to the analytic focus on the system, the network of interacting or interdependent actors, and the social patterns and outcomes that emerge and are reproduced through the generativity of the action basis. (Fararo and Skvoretz 2002: 296)

Methodologically, this orientation seeks mathematical expression of generative process models and where there is considerable complexity in the concatenations of processes, it turns to simulation as a tool to investigate the consequences implied by the assumptions that specify the models. This step highlights the importance of the emergence of computational sociology in the last decades of the twentieth century (Hummon and Fararo 1995b). Recent conceptual and theoretical statements relating to generative process models tend to confirm the judgment that a significant advance in the methodology of social theory is underway (Cederman 2005; Cherkaoui 2005; Manzo 2007).

REFERENCES

Anderson, J.R. 1993. *Rules of the Mind*. Hillsdale: Lawrence Erlbaum Associates.

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

Axten, N. and T.J. Fararo. 1977. "The information processing representation of institutionalized social action," in P. Krishnan (ed.), *Mathematical Models of Sociology. Sociological Review Monograph* 24: 35–77. Keele, UK.

Bainbridge, W., E. Brent, K. Carley, D. Heise, M. Macy, B. Markovsky and J. Skvoretz. 1994. "Artificial social intelligence," *Annual Review of Sociology* 20: 407–36.

Berger, J. and M. Zelditch, Jr. (eds.) 1985. *Status, Rewards and Influence*. San Francisco: Jossey-Bass.

2002. *New Directions in Contemporary Sociological Theory*. Lanham: Rowman & Littlefield.

Berger, J., M. Zelditch, Jr. and B. Anderson. 1972. "Introduction," in J. Berger, M. Zelditch, Jr. and B. Anderson (eds.), *Sociological Theories in Progress*, vol. II. Boston: Houghton Mifflin.

Berger, P. and T. Luckmann. 1966. *The Social Construction of Reality*. New York: Doubleday.

Braithwaite, R.B. 1953. *Scientific Explanation*. Cambridge University Press.

Cederman, L. 2005. "Computational models of social forms: advancing generative process theory," *American Journal of Sociology* 110(4): 864–93.
Chase, I.D. 1980. "Social process and hierarchy formation in small groups: a comparative perspective," *American Sociological Review* 45: 905–24.
Chase, I.D. and W.B. Lindquist. 2009. "Dominance hierarchies," in P. Hedström and P. Bearman (eds.), *The Oxford Handbook of Analytical Sociology*. Oxford University Press.
Cherkaoui, M. 2005. *Invisible Codes: Essays on Generative Mechanisms*. Oxford: Bardwell Press.
Cohen, B.P. 1963. *Conflict and Conformity: a Probability Model and its Application*. Cambridge, MA: MIT Press.
Davis, A., B.B. Gardner and M.R. Gardner. 1941. *Deep South: a Social Anthropological Study of Caste and Class*. University of Chicago Press.
Evans, J.B. 1988. *Structures of Discrete Event Simulation*. Englewood Cliffs: Prentice Hall.
Fararo, T.J. 1969. "The nature of mathematical sociology," *Social Research* 36(1): 75–92.
1972. "Dynamics of status equilibration," in J. Berger, M. Zeldtich, Jr. and B. Anderson (eds.), *Sociological Theories in Progress*, vol. II. Boston: Houghton Mifflin, Chapter 9.
1973. *Mathematical Sociology*. New York: Wiley.
1989. *The Meaning of General Theoretical Sociology*. New York: Cambridge University Press.
2001. *Social Action Systems*. Westport: Praeger.
Fararo, T.J. and C. Butts. 1999. "Advances in generative structuralism: structured agency and multilevel dynamics," *Journal of Mathematical Sociology* 24(1): 1–65.
Fararo, T.J. and N.P. Hummon. 1994. "Discrete event simulation and theoretical models in sociology," *Advances in Group Processes* 11: 25–66.
Fararo, T.J. and K. Kosaka. 2003. *Generating Images of Stratification: a Formal Theory*. Dordrecht: Springer.
Fararo, T.J. and J. Skvoretz. 1984. "Institutions as production systems," *Journal of Mathematical Sociology* 10: 117–81.
1986. "Action and institution, network and function: the cybernetic concept of social structure," *Sociological Forum* 1: 219–50.
2002. "Theoretical integration and generative structuralism," in J. Berger and M. Zelditch, Jr. (eds.), *New Directions in Contemporary Sociological Theory*. Lanham: Rowman & Littlefield, 295–316.
Fararo, T.J., J. Skvoretz and K. Kosaka. 1994. "Advances in E-state structuralism: further studies in dominance structure formation," *Social Networks* 16(3): 233–65.
Heise, D.R. 1989. "Modeling event structures," *Journal of Mathematical Sociology* 14: 139–69.
Homans, G.C. 1950. *The Human Group*. New York: Harcourt Brace & World.
1958. "Social behavior as exchange," *American Journal of Sociology* 63: 597–606.
1974 [1961]. *Social Behavior: its Elementary Forms*, rev. edn. New York: Harcourt Brace Jovanovich.

Hummon, N.P. and T.J. Fararo. 1995a. "Actors and networks as objects," *Social Networks* 17: 1–26.

1995b. "The emergence of computational sociology," *Journal of Mathematical Sociology* 20(2–3): 79–87.

Manzo, G. 2007. "Variables, mechanisms and simulations: can the three methods be synthesized?" *Revue française de sociologie* 48, Supplement: 35–71.

McClelland, K. and T.J. Fararo. (eds.) 2006. *Purpose, Meaning and Action: Control Systems Theories in Sociology*. New York: Palgrave Macmillan.

McPhail, C., D.S. Schweingruber and A. Ceobanu. 2006. "Purposive collective action," in K. McClelland and T.J. Fararo (eds.), *Purpose, Meaning and Action: Control Systems Theories in Sociology*. New York: Palgrave Macmillan, Chapter 3.

Margenau, H. 1950. *The Nature of Physical Reality*. New York: McGraw-Hill.

Miller, G.A., E. Galanter and K. Pribram. 1960. *Plans and the Structure of Behavior*. New York: Holt.

Nadel, S.F. 1951. *Foundations of Social Anthropology*. Glencoe, IL: Free Press.

Newell, A. and H.A. Simon. 1972. *Human Problem Solving*. Englewood Cliffs: Prentice Hall.

Parsons, T. 1937. *The Structure of Social Action*. New York: McGraw-Hill.

1951. *The Social System*. New York: Free Press.

1960. "Pattern variables revisited," *American Sociological Review* 25: 467–83.

Powers, W.T. 1973. *Behavior: the Control of Perception*. Chicago: Aldine.

Simon, H.A. 1952. "A formal theory of interaction in social groups," *American Sociological Review* 17: 202–12.

Skvoretz, J. and T.J. Fararo. 1988. "Dynamics of the formation of stable dominance structures," in M. Webster and M. Foschi (eds.), *Status Generalization*. Palo Alto: Stanford University Press, Chapter 17.

1995. "The Evolution of Systems of Social Interaction." *Current Perspectives in Social Theory* 15: 275–299. Greenwich, CT: JAI Press.

1996. "Status and participation in task groups: a dynamic network model," *American Journal of Sociology* 101: 1366–414.

Skvoretz, J., M. Webster and J. Whitmeyer. 1999. "Status orders in task discussion groups," *Advances in Group Processes* 16: 199–218.

Veblen, T. 2007 [1899]. *The Theory of the Leisure Class*. Oxford University Press.

Whitehead, A.N. 1925. *Science and the Modern World*. New York: Macmillan.

Wooldridge, M. 2002. *An Introduction to Multi-Agent Systems*. New York: Wiley.

6 Singular mechanisms and Bayesian narratives

Peter Abell

Introduction

I shall interpret the social sciences, including sociology, as a quest for causal mechanisms. It is not clear, though, how we should handle situations where it proves difficult or even impossible to find repeated observations of the events which are purportedly in a causal relationship. Historically focused case studies often seem to involve unique or low frequency events, though conjectured causal connections are routinely deployed.[1]

The study of causal mechanisms which account for (or show how) patterns of co-variation are established (generated) by human activity lies at the foundations of analytical sociology. Such mechanisms may be postulated at a latent theoretical level or be subject to direct observation. The boundary between observational and theoretical terms is often flexible depending upon improved observational and measurement techniques. The manner of causal inference, where events are repeated, is well understood, though any inference is always provisional subject to as yet unexplored exogeneity tests, but when it comes to relative rare events the story is much less clear. How can, if at all, causal inferences be safely made in such circumstances?

The chapter develops as follows. First, I briefly review the way in which causal inferences are vouchsafed in repeated observational (i.e. large N) studies. Second, the idea of singular causality is developed based upon the concept of human action. This licenses the notion that sequences of causally connected actions become the focus of attention and thus, third, the concept of narrative takes center stage. Fourth, the concept of causality and narratives are linked and related to the

[1] It proves rather difficult to define what a case study is (see for instance Ragin and Becker 1992). Repeated observations on a single unit of analysis can of course generate a (univariate or multivariate) time series which is open to (Granger) causal analysis. It is not usual to regard a time series as a case study, though a case may incorporate a time series. A case study is likely to involve repeated observations on a single unit, but registering a variety of concepts/variables at different points in time.

historian's concept of colligation whereby macro concepts of causality are decomposed into finer grained ones. Fifth, the concept of Bayesian narrative is introduced whereby causal links are studied not in terms of detecting a generalization, but by the compilation of evidence for and against a causal link. Sixth, inferential procedures are briefly considered and then the chapter concludes.

Causal inference in large-N observational studies

It is standard practice when engaging in large-N studies to adopt, at least implicitly, a Humean, constant conjunction, interpretation of any causal link, C (cause) → E (effect). When suitable controls are in place and C is temporally prior to E, repeated cases of C leading to E, along with cases where not-C leads to not-E (i.e. comparative method) need to be observed. We all know that constant conjunction (in practice, statistical co-variation) neither implies nor is implied by causality but with careful functional specification, controls and explicit assumptions, co-variation and causality can be effectively reconciled (Pearl 2000; Spirtes *et al*. 2000). All causal inferences are, however, provisional and therefore vulnerable to revision in terms of, as yet, unconducted endogeneity tests which might establish a spurious relationship between C and E. Randomized experiments are more propitiously positioned in this respect but such research designs are rarely feasible in the social sciences.

Technical details apart, the underlying logic of large-N observational studies is transparent; both comparative method (i.e. comparing units of analysis or cases) and generalization (preferably perhaps nomothetic) are prerequisites of any causal explanation (and, thus, prediction). This ordering of the "trinity" – explanation necessitating prior comparison and generalization – is one possible way of understanding the epistemological assumptions of Positivism, ensconced in both the standard hypothetico-deductive (covering-law) and inductive-probabilistic models of explanation. In the context of the latter, the generalization will take the form of an increased probability of E given C compared with E in the absence of C.

I have argued elsewhere (Abell 2001, 2004) that for causal inference to prove at all possible in a low-N situation a reordering of the (Positivist?) trinity (explanation, comparative method, generalization) is necessary.[2] Singular causal explanation (i.e. making the inference C causes E in a

[2] In the light of footnote 1 I use the term small N to mean the number of major units of analysis not the number of observations.

particular case, without the benefit of generalization) must be rendered logically prior to both comparison and any subsequent limited generalization to a few cases. So then we may cogently ask, by comparing a few causal explanations across cases, whether or not an explanation can be generalized. Contrast the two questions "how generalizable is a (singular) causal explanation?" and "is there a known generalization which licenses a causal explanation?"

The reordering I am advocating is not merely an indirect restatement of the precepts of the classical inductive probabilistic form of explanation. There, cases of C associated with E (and not-C with not-E) are assembled in order to determine whether or not a (probabilistic) generalization occurs such that we might conceive of an explanatory causal link operating between them and, indeed, the likelihood of the truth of the appropriate counterfactual. On the contrary, the reading I am searching for would inductively assemble cases (i.e. make use of comparative method, if any comparable cases exist) in order to ask whether or not (or the degree to which) a known singular causal explanation is open to generalization.

Singular causality

Much, thus, depends upon finding a satisfactory conception of singular causality and specifying the conditions under which it can be observed. In this regard I have adapted Von Wright's (1971) conception of the practical syllogism creating the contingent syllogism (Abell 1987, 2004, 2009).[3] Individual and collective actions may, I have argued, be regarded as singular causes of further actions. If this is so, sequences of causally connected actions become the focus of enquiry (i.e. Narratives). Insofar as we have evidence for both the actions and their causal connection there is a sense in which we can ask whether, "beyond all reasonable doubt" (a concept to be explored below), a singular causal relationship exists. There is no question of discovering the cause from a comparative pattern of (Humean) constant conjunction of the actions, though intentional explanations can, of course, incidentally depend upon generalized (perhaps nomic) connections between real world events. Clearly, they make actions possible in the physical world. If I try to do X, under the belief that X will eventuate in a physical world event E, and X and E are not causally linked then I shall be disappointed. I distinguish between a necessary component of and the possible grounds for a causal explanation (Danto 1985). Physical

[3] Von Wright does not, I think, promote a singular notion of causality.

causality provides grounds for narrative causality but the latter has its own logical structure.

Social scientists are often interested in causal relations of the form: "α doing X_1 causes, shall we say, β to do something else" (i.e. a form of "social" interaction).[4] The issue, then, is how this sort of causality can be conceived in singular terms.

Singular, action-driven causality enables an unorthodox interpretation of counterfactual evidence. Suppose we wish to explore whether "β did X_2" because "α did X_1" so, disclosing that α doing X_1 caused β to do X_2.[5] We may be in possession of counterfactual testimony licensing the statement that "β would not have done X_2 if α had not done X_1." Indeed, β would have done something else. If reliable evidence/testimony of this sort, despite its controversial nature, is permitted an explanation of a causal effect is at hand, without any resort to comparison or generalization (i.e. across αs and βs). The most obvious sorts of evidence, in the analysis of contemporary cases, are statements from the actors concerned, either given spontaneously or solicited. Where such evidence is not available (e.g. with historical case studies) then evidence will be more indirect – sometimes very indirect. It may take the form of recorded testimony of contemporary protagonists or observers or of historians themselves. In this latter sense historians search for varieties of indirect evidence to support a proposition describing interactions such as "β did X_2 because α did X_1." The important point here is that in sifting the balance of evidence they will not characteristically see it as appropriate to search for matching interactions but rather to assemble as much evidence as they can muster for the particular interaction under scrutiny. It is however imperative, in my view, not to interpret, as many advocates of case studies do, this search for detailed evidence as warranting an essentially deterministic view of the exercise. The relationship between the assembled evidence for a causal link and its "truth" must still be conceived in probabilistic terms.[6]

[4] The *link* between "α doing X_1" and "β doing something else" may or may not be the intention of α. But for the sake of clarity I shall ignore this complication. All that narrative analysis (below) requires is that some state produced by α's action has an impact upon β's reasoning and eventual action.

[5] The "because" here is ambiguous with respect to reasoned (by β) necessity or sufficiency.

[6] As Goldthorpe (2000) has pointed out there is an unwarranted tendency amongst those who advocate case studies to assume that because cases are studied in great detail they are observed without error (i.e. deterministically). As I shall show below Bayesian narratives are, as the name might suggest, essentially probabilistic though the probabilities are based upon degrees of subjective belief rather than frequency or aleatory considerations.

The role of the counterfactual in this reasoning does not derive from comparison but is internal and derived from the particular case. If evidence can be found supporting the counterfactual that "β would not have done X_2 if α had not done X_1" then this enhances the conclusion that the doing of X_1 caused the doing of X_2. Such evidence absent the causal conclusion is weakened but not vitiated since other sorts of evidence for the causal connection may be available.

I have tried in foregoing paragraphs, and more thoroughly elsewhere (Abell 2009), to establish a distinctive role for single cases in causal analyses. My argument depends, first, upon an acceptance of a singular notion of (action-driven) causality but, second, also upon a structuring of actions within cases in a particular manner whereby paths of interaction can be traced out. That is to say upon a conception of narrative.

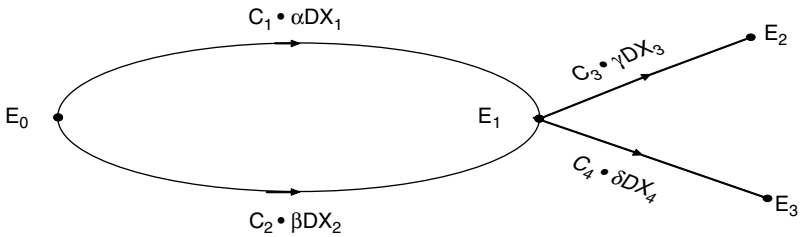
Narratives

I now turn to a brief exposition of the concept of narrative. I shall, however, only offer an informal presentation of the ideas (for a fuller presentation see Abell 1987, 1993). For a similar approach see also Heise (1989).

Essentially a narrative comprises a time-ordered a-cyclic multi-arc di-graph where the *nodes* (vertices) depict states of the world and the *arcs* the actions which transform these states. Two or more arcs incident *into* a node implies that they (i.e. the actions) are *conjointly* sufficient for the occurrence of the node. Two or more arcs incident out of a node imply that the node provides a condition for each action. In Figure 6.1 the notation – D – is used to depict an action generating a simple narrative structure where, first, the actions αDX_1 (α does X_1) and βDX_2 jointly change the world from E_0 to E_1. Subsequently, E_1 provides a condition for the actions by both γ and δ in realizing, respectively, E_2 and E_3. In each case the actions are also prompted by conditions C.

If $E_2 = E_3$ then γDX_3 and δDX_4 are jointly sufficient for this state of affairs. Often, of course, it may be that $C_1 = C_2 = C_3 = C_4$ – the common condition for the actions constituting the narrative. Furthermore, for practical purposes the structure in Figure 6.1 may be depicted as the di-graph given in Figure 6.2 where nodes depict actions (including conditions C) and responses to states of the world are dropped, giving an "action skeleton."[7] It is often convenient to think in terms of such

[7] The action skeleton is derived in the following manner. By virtue of the contingent practical syllogism we may say: "in a situation perceived by α as C, α intended X_1 and believed that doing a_1 would procure X_1. Therefore α did a_1. In situation a_1 and X_1, β



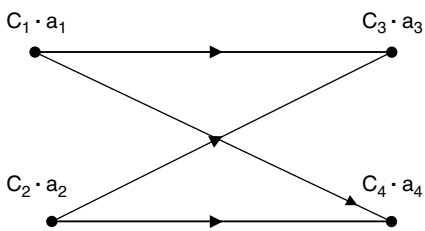
α, β, γ, δ – Actors

C_1, C_2, C_3, C_4 – Conditions

DX_2 – Action Performed

E_0, E_1, E_2, E_3 – States

Figure 6.1 A narrative.



$a_1 = \alpha DX_1$
$a_2 = \beta DX_2$
$a_3 = \gamma DX_3$
$a_4 = \delta DX_4$

Figure 6.2 An action skeleton.

a di-graph, and I shall tend to do so in what follows using the simple notation a_1, a_2, etc.

A narrative may be regarded as either an explanation of how the terminal node(s) is (are) causally generated or of how the initial node(s) is transformed to the final node(s) (two alternative explicanda). In the former context, the further into the past the narrative is constructed, the more of the history of its causal generation (mechanisms) is rendered explicit.

intended X_2 and believed that doing a_2 would procure X_2. This is then shortened to α did a_1 caused β to do a_2."

Narratives and causal inference

Many historians propose that causal links in a narrative should be decomposed until the finer grained causal links, so produced, can be understood in terms of "laws which we derive from our everyday experience" (Roberts 1996). Roberts calls this decomposition micro-colligation and since the decomposed causality explicitly depends upon laws, the explanation now appears to fit the classical hypothetico-deductive paradigm. As a consequence this interpretation of micro-colligation is consistent with the classical ordering of the explanation/comparison/generalization trinity.

It does not always seem tenable, however, to argue that fine-grained causal connections are characteristically examples of causal laws. Rather what the decomposition, in my view, should achieve is a narrative formulated in terms of sequential actions which bear a singular causal interpretation.

Colligation and Bayesian narratives

Specifying the nodes of a narrative, comprising a chronology of temporally ordered events (actions), often proves relatively uncontroversial though the decision to do something and its actual accomplishment may be separated in time which can complicate any causal ordering. Historians, nevertheless, only infrequently dispute the occurrence and dating of events or actions. It is the insertion of causal links, between pairs of events (actions), thus creating a narrative, which more often than not proves problematic. In what follows I am going to assume that the chronology of actions is given and, thus, the issue is to explore how causal links can be interposed between them, but the techniques proposed could equally be applied to the chronology of events/actions.

The basic building blocks of a narrative (expressed as an action skeleton as in Figure 6.2) take the form:

Action a_1 by α causes action a_2 by β,

where α and β may be individual or collective and may or may not be distinct. In general the question we wish to pose is, how do the odds that the causal link is present alter, given various items of evidence which have been assembled? The items of evidence, some favoring and some opposing the presence of a causal link, may be either independent of each other or dependent, conditional upon the link (e.g. testimony of two or more independent observers as opposed to colluding observers). The objective

of causal analysis is to find a way of combining the separate items of evidence in a consistent manner to demonstrate that the odds of the presence of the link is established "beyond all reasonable doubt."

Let us call the hypothesis that a causal link does, in this case, pertain A and its denial ¬A. So, we now ask what is the evidence, in this particular case, for the truth of A? Assume, initially, one item of evidence only, b (e.g. a report by β that she did a_2 because α did a_1),[8] in substantiating P(A | b), the probability that a causal link is present given that β reports that "she did a_2 because α did a_1."

So, using Bayes' law we have:

$$P(b) \cdot P(A \mid b) = P(A) \cdot P(b \mid A) \quad (1)$$
$$\text{and } P(b) \cdot P(\neg A \mid b) = P(\neg A) \cdot P(b \mid \neg A). \quad (2)$$

Where P(A) is the probability of A, P(b) is the probability of b, P(b | A) is the probability of b given A and P(b | ¬A) is the probability of b given ¬A.

$$\text{Thus, } \frac{P(A \mid b)}{P(\neg A \mid b)} = \frac{P(A) \bullet P(b \mid A)}{P(\neg A) \bullet P(b \mid \neg A)} \quad (3)$$

$$\text{So, Odds } ((A : \neg A) \mid b) = \text{Odds } (A : \neg A) \cdot L_b \quad (4)$$

where, $L_b = P(b \mid A) / P(b \mid \neg A)$ is the likelihood ratio of the evidence b on A and ¬A.

$$\text{So, } \log L_b = \text{Log Odds } ((A: \neg A) \mid b) - \text{Log Odds } (A : \neg A). \quad (5)$$

Log L_b (Good 1983) gives a measure of how the odds of A as against ¬A alters as a consequence of the evidence b. It provides a measure of the evidential support that b gives for the existence of the causal link A by changing the prior odds to the posterior odds. Expressing equation (4) in logarithmic terms (equation (5)) is clearly optional; it merely allows an interpretation of the change in odds in terms of a difference and locates no evidential impact at zero.

[8] The evidence b might be more indirect than the testimony of a participant. Even such testimony must be regarded as only probabilistically relevant to the presence of the causal link (A) – participants may deceive etc. The only requirement is that P (A | b) > 0 or P (¬A| b) > 0; that is the evidence has some relevance one way or the other for the causal inference. The precise interpretation of P (b | A) will depend upon the nature of the evidence b. In the special case of a report by a participant (β) that "I did a_2 because of a_1" then P(b | A) is the probability that β gives such a report given the truth of A. Similarly, P(b | ¬A) is the probability of the report, given A, is false. Of course such reports will when examined more closely depend upon an estimate (by the analyst) of the understanding β has of her own motivation and her propensity to report truthfully. The probabilities of both β's understanding and truth-telling propensity (conditionally dependent on A and ¬A) may, at this finer grained analysis, be interpreted as states intervening in the inference between A and b. I show below how to deal with such states.

Log $L_b = 0$ (i.e. $L_b = 1$) if b has no impact on the odds; ($P(b \mid A) = P(b \mid \neg A)$),

Log $L_b > 0$ (i.e. $L_b > 1$) if b has a positive impact on the odds; ($P(b \mid A) > P(b \mid \neg A)$),

Log $L_b < 0$ (i.e. $L_b < 1$) if b has a negative impact on the odds; ($P(b \mid A) < P(b \mid \neg A)$).

In practice there may be two or more items of evidence which bear upon the likelihood of the truth or falsity of hypothesis A. Start by assuming there are two items, b_1 and b_2, mutually independent conditional upon A and ¬ A (e.g. two entirely independent reports).

Then we have,

$$P(A \mid b_1,b_2) \cdot P(b_1) \cdot P(b_2) = P(A) \cdot P(b_1 \mid A) \cdot P(b_2 \mid A)$$
$$\text{and } P(\neg A \mid b_1,b_2) \cdot P(b_1) \cdot P(b_2) = P(\neg A) \cdot P(b_1 \mid \neg A) \cdot P(b_2 \mid \neg A). \quad (6)$$

Dividing the first equation by the second:

$$\frac{P(A \mid b_1,b_2) \bullet P(b_1) \bullet P(b_2)}{P(\neg A \mid b_1,b_2) \bullet P(b_1) \bullet P(b_2)} = \frac{P(A) \bullet P(b_1 \mid A) \bullet P(b_2 \mid A)}{P(\neg A) \bullet P(b_1 \mid \neg A) \bullet P(b_2 \mid \neg A)}$$

Cancelling P(b1) and P(b2) gives:

$$\frac{\text{Odds } ((A : \neg A) \mid b_1, b_2)}{\text{Odds } (A : \neg A)} = \frac{P(b_1 \mid A) \bullet P(b_2 \mid A)}{P(b_1 \mid \neg A) \bullet P(b_2 \mid \neg A)} = L_{b1} \bullet L_{b2} \quad (7)$$

So, Log Odds $((A: \neg A) \mid b_1, b_2)$ – Log Odds $(A : \neg A)$ = Log L_{b1} + Log L_{b2} . (8)

$$\text{Let, } L_B = L_{b1} \cdot L_{b2} \quad (9)$$

where, L_B is the likelihood ratio of b_1 and b_2 conditional upon A and ¬A.

So, in general, for n conditionally independent items of evidence we will then have:

$$L_B = L_{b1} \cdot L_{b2} \cdot \cdot L_{bn} \quad (10)$$
$$\text{and Log } L_B = \Sigma_n \log b_n. \quad (11)$$

Equation (11) details how all the items of conditionally independent evidence combine in support of or against the hypothesis A (alternatively, support of or against hypothesis ¬A). As we shall see below L_B gives an estimate of how the combined (conditionally independent on A and ¬A) items of evidence alter the log-odds of A as against ¬A; that is to say, the log-odds for the existence of a causal link between the actions/events specified in the chronology.

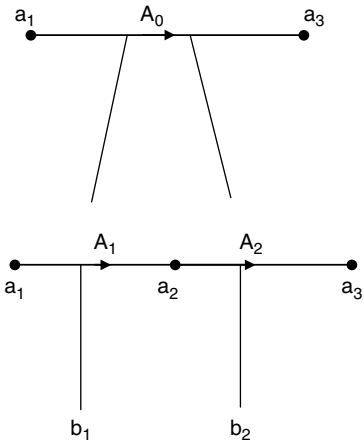


Figure 6.3 Colligation.

Now assume that b_1 and b_2 are not independent conditional upon A and ¬A (e.g. two partially collaborative reports).

So,[9]

$$P(A\ |b_1,b_2) \cdot P(b_1,b_2) = P(A) \cdot P(b_1|A) \cdot P(b_2|A,,b_1)$$
$$P(\neg A|b_1,b_2) \cdot P(b_1,b_2) = P(\neg A) \cdot P\ (b_1|\ \neg A) \cdot P(b_2|\neg A\ ,b_{11}). \qquad (12)$$

Dividing and taking logs,

$$\text{Log Odds}\ ((A: \neg A)|\ b_1,b_2) - \text{Log Odds}\ (A: \neg A) = \text{Log}\ L_{b1} + \text{Log}\ L_{b2/b1} \qquad (13)$$

where $L_{b2/b1}$ is the likelihood ratio of b_2 given b_1 conditional on A and ¬ A. Thus, with multiple conditionally dependent items of evidence we have equations similar to (8) and (9) but reflecting the pattern of conditional dependence amongst the items of evidence.

Comparing equations (13) with (8) demonstrates that the conditional dependence of items of evidence does not materially alter the picture. The change in log odds, given two items of evidence, is the sum of the appropriate log likelihood ratios in both cases.

Clearly, the analysis may be extended to any number of items of conditionally dependent evidence.

In order to gain some analytical insight consider, first, the single decomposition depicted in Figure 6.3 where hypothesis A_0 (a causal link exists between a_1 and a_3) is decomposed (colligated) into two

[9] Note that $P(b_1,b_2) = P(b_1)\ P(b_1|b_2) = P(b_2)\ P(b_2|b_1)$ – conditional dependence is symmetric.

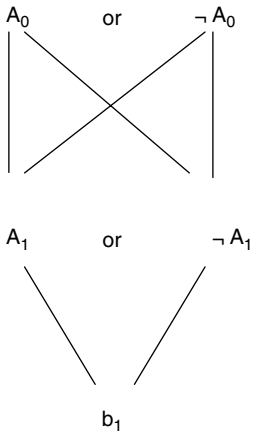


Figure 6.4 The decomposition of hypotheses A_0 and $\neg A_0$.

hypotheses labeled A_1 (a causal link exists between and a_1 and a_2) and A_2 (a causal link exists between a_2 and a_3).

Let us further assume we have two items of evidence, b_1 and b_2, one each respectively for the hypotheses A_1 and A_2.

Now assume that:

1. b_1 and b_2 are independent conditional upon A_0 and $\neg A_0$.
2. b_1 is independent of A_0 and $\neg A_0$ conditional upon A_1 and $\neg A_1$. This is implied by the lack of a direct link in Figure 6.3 running between A_0 and b_1.
3. b_2 is independent of A_0 and $\neg A_0$ conditional upon A_2 and $\neg A_2$. Again this is implied by the lack of a direct link running between A_0 and b_2.
4. A_1 and A_2 are independent conditional upon A_0 and $\neg A_0$.

Then following Figure 6.4 we may write:

$$P(b_1 \mid A_0) = P(A_1 \mid A_0) \cdot P(b_1 \mid A_1) + P(\neg A_1 \mid A_0) \cdot P(b_1 \mid \neg A_1)$$
$$P(b_1 \mid \neg A_0) = P(A_1 \mid \neg A_0) \cdot P(b_1 \mid A_1) + P(\neg A_1 \mid \neg A_0) \cdot P(b_1 \mid \neg A_1). \quad (14)$$

$$\text{So,} \quad L_{b1} = \frac{P(A_1 \mid A_0) \cdot P(b_1 \mid A_1) + P(\neg A_1 \mid A_0) \cdot P(b_1 \mid \neg A_1)}{P(A_1 \mid \neg A_0) \cdot P(b_1 \mid A_1) + P(\neg A_1 \mid \neg A_0) \cdot P(b_1 \mid \neg A_1)} \quad (15)$$

Similarly,

$$L_{b2} = \frac{P(A_2 \mid A_0) \cdot P(b_2 \mid A_2) + P(\neg A_2 \mid A_0) \cdot P(b_2 \mid \neg A_2)}{P(A_2 \mid A_0) \cdot P(b_2 \mid A_2) + P(A_2 \mid A) \cdot P(b_2 \mid A_2)} \quad (16)$$

And $L_B = L_{b1} \cdot L_{b2}$. (17)

Generalizing to n items of evidence,

$$L_B = L_{b1} \cdot L_{b2} \cdot \ldots\ldots\ldots \cdot L_{bn} \quad (18)$$

Thus, with n items of evidence all of which are pairwise independent conditional upon the ultimate hypothesis under investigation (A_0 and $\neg A_0$) the aggregate probative force of the evidence is, as before, computed by the multiplication of the odds ratios of each item. Since, however, there are intervening hypothesis (A_1 and A_2) lying between the items of evidence and A_0 and $\neg A_0$ the odds ratios are themselves computed from constituent components (equations (15) and (16)).

Relaxing assumption (1) so that the items of evidence $b_1$and b_2 are now dependent conditional on A_0 and $\neg A_0$

$$L_B = L_{b1} \cdot L_{b2/b1} = L_{b2} \cdot L_{b1/b2} \quad (19)$$

where

$$L_B = P(b_2 \mid A_0, b_1) / P(b_2 \mid \neg A_0, b_1) = \text{Odds}((A_0 : \neg A_0) \mid b_1, b_2) / \text{Odds}(A_0 : \neg A_0). \quad (20)$$

Generalizing to n items of evidence,

$$L_B = L_{b1} \cdot L_{b2/b1} \cdot L_{b3/b2 \cdot b1} \ldots\ldots\ldots\ldots \cdot L_{bn/b1 \cdot b2 \ldots\ldots \cdot b.(n-1)} \quad (21)$$

To make inferential use of these sorts of equations estimates of the various likelihood ratios are needed. We now turn in this direction.

Inferential procedures

Given a conjectured causal link in a chronology (A_0 above) which is then decomposed into finer grained causal links, ultimately, each of which are linked to item(s) of evidence (b_1, b_2 …. above), we wish to attach a value to the posterior odds given the evidence. The value, in the absence of repeated observation (frequencies), will have to be derived from the "degrees of belief" of the analyst (Schum 1994). The analyst, in turn, may rely upon the estimates of informed respondents, for example, knowledgeable historians. There is of course a lengthy tradition in sociology of placing careful reliance upon the testimony of experts or "key informants." March *et al.* (1991), in a provocatively entitled paper "Learning from samples of one or fewer," write as follows:

> Theories of historical inference tend to emphasise pooling of observations. Pooling over observers appears to have advantages in some common situations but in the absence of a clearer formulation of the gains and losses involved, it is hard to specify the precise conditions favouring one strategy or the other.

Bayesian narratives, I conjecture, might offer some help in this respect.

It is difficult to see, in the absence of comparative cases, how this conjecture can be dispensed with. Of course the decomposition of link A_0 may lead to finer grained links, each of which can be repeatedly observed when the appropriately derived correlations between the states connected by the causal link can be inserted into the chronology.

Since an estimate of the posterior odds, conditional upon the evidence, is our ultimate objective the analyst could of course ask any key respondent to make a stab at its value. We could perhaps, following Simon, even average the values given by a number of such respondents. Giving such estimates, however, would prove both difficult for the respondents to achieve with any reliability and would not allow for comparative estimates of the impact of individual items of evidence upon the posterior odds. Rather, respondents should be offered a framework in which they can assemble their estimates of the odds from the evidence, whilst checking upon the consistency of their reasoning.

The general inferential structure from the forgoing analysis takes the form:

$$\text{Odds}\,((A{:}\,\neg A) \mid b_1, b_2, \ldots b_n) = L_B = L_{b1} \cdot L_{b2/b1} \ldots \cdot L_{bk/b1\cdots bk-1} \ldots \cdot L_{bn/b1\cdots bn-1} \cdot \text{Odds}(A{:}\neg A) \tag{22}$$

In order to estimate the posterior Odds $((A{:}\neg A) \mid b_1, b_2 \ldots\ldots, b_n)$ from L_B we still require an estimate of the prior odds. I shall return to that issue presently.

Expert informants may make estimates of:

1. the constituent likelihoods in the appropriate equation for L_B;
2. the global likelihood L_B.

The analyst can then search for consistency between the these estimates before accepting the estimate of the global L_B. Alternatively, the likelihoods could be estimated as though the items of evidence are severally conditionally independent on A_0 and $\neg A_0$ and any diversity between their product and the global estimate of L_B could then be ascribed to conditional dependence amongst the items of evidence. In general, the analyst will discount those estimates by key informants whom are neither consistent nor embrace a wide range of evidential items.

Since the objective is to estimate odds $((A \cdot \neg A) \mid b_1, b_2 \ldots, b_n)$ we still need an estimate if the prior odds of A_0 to $\neg A_0$. There are probably two reasonable approaches to this estimation:

1. assume the odds are 1, (i.e. in the absence of any evidence A_0 and $\neg A_0$ are equally probable);
2. ask the key informant for an estimate (see Schum (1994) for a legal analysis).

Then in either case the posterior odds can be computed from the prior odds and the appropriate likelihood ratios.

Finally, do the odds so computed permit us to infer "beyond all reasonable doubt" that either A_0 or $\neg A_0$ is true? As with any conception about causality based upon frequencies one can impose more or less demanding criteria in respect of significance.

The values of $P(A_0)$ and $P(\neg A_0)$ are given by:

$$P(A_0) = X/(1+X) \tag{23}$$

$$P(\neg A_0) = 1/(1+X) \tag{24}$$

where X :1 is the value of the odds that A_0 is true.

By setting the prior odds to unity we assume, in the absence of any evidence, that $P(A_0)$ and $P(\neg A_0)$ are identical in value and therefore both equal to 0.5. In the presence of evidence we might reasonably require that when the posterior odds are 100:1 then A_0 is true "beyond all reasonable doubt" (a causal connection exists in Fig. 6 between a_1 and a_3). Furthermore, when the odds are 1:100 then $\neg A_0$ is true "beyond all reasonable doubt" (there is no causal connection between a_1 and a_3). Thus, the range of change in odds is four log units.

The reader might feel rather disconcerted about the subjective nature of these various estimates but my claim is that, by being explicit about the inferences and incorporating the internal consistency checks, a systematic inferential procedure is at hand which permits causal inference without inter-case comparison.

Conclusion

Those who advocate the precepts of an analytical sociology assume that we can explore the nature of causal mechanisms involving human actions (and forbearances) which drive changes in the social world. The method of achieving this objective when events are repeatable such that statistical inferences can be realized are well understood. Where, however, events are rare and, in the absence of established causal generalizations, causal inference is more problematic, I have outlined an alternative method of causal inference appropriate to such circumstances which I call the method of Bayesian narratives.

REFERENCES

Abell, P. 1987. *The Syntax of Social Life: The Theory and Method of Comparative Narratives*. Oxford University Press.

1993. "Some aspects of narrative method," *Journal of Mathematical Sociology* 18: 93–134.

2001. "Causality and low-frequency complex events: the role of comparative narratives," *Sociological Methods and Research* 30: 57–80.

2004. "Narrative explanation: an alternative to variable centered explanation," *Annual Review of Sociology*. 30: 287–310.

2009. "A case for cases," *Sociological Methods and Research* 32: 1–33.

Danto, A.C. 1985. *Narration and Knowledge*. New York: Columbia University Press.

Goldthorpe, J.H. 2000. *On Sociology*. Oxford University Press.

Good, I.J. 1983. *Good Thinking: the Foundations of Probability and its Applications*. Minneapolis: University of Minnesota Press.

Heise, D. 1989. "Modelling event structures," *Journal of Mathematical Sociology* 14: 39–169.

March, J.G., L.S. Sproull and M. Tamuz. 1991. "Learning from samples of one or fewer," *Organization Science* 2: 1–13.

Pearl, J. 2000. *Causality*. Cambridge University Press.

Ragin, C.C. and H.S. Becker (eds.) 1992. *What is a Case? The Foundations of Social Inquiry*. Cambridge University Press.

Roberts, C. 1996. *The Logic of Historical Explanation*. Philadelphia: Pennsylvania University Press.

Schum, D.A. 1994. *The Evidential Foundations of Probabilistic Reasoning*. Evanston: Northwestern University Press.

Spirtes, P., C. Glymour and R.S. Scheines. 2000. *Causation, Prediction and Search*. New York: Springer-Verlag.

Von Wright, G.H. 1971. *Explanation and Understanding*. London: Routledge and Kegan Paul.

7 The logic of mechanistic explanations in the social sciences

Michael Schmid

Statement of the problem

This chapter draws together some philosophical (Bunge 2004; Little 1998) as well as sociological arguments (Balog 2006; Hedström 2005; Manicas 2006) in favor of an explanatory and realistic research program for the social sciences. As I shall show, such a program presupposes that social scientific explanations are to be couched in the form of microfoundational multi-level explanations of macroscopic states of affair with reference to a substantive theory of individual action.

The logic of scientific explanations

The background to my subject was described by Carl Hempel when he proposed that the social sciences call only those explanations successful that can satisfy a series of "conditions of adequacy" (Hempel 1965: 247ff.). Among these are that the explanandum can be logically deduced from the explanans; moreover, that the explanans should contain at least one deductively necessary nomological proposition or "law"; further, that the propositions of an explanatory argument should be empirically confirmable; and finally, that the propositions of the explanans must be true.

Against this explanatory scheme quite different criticisms were brought forth that increasingly denied the capacity of the social sciences to furnish explanations of the kind that Hempel defended (Bayertz 1980). Among the most important criticisms was that the statement of general laws is not a necessary condition for explanatory success (Scriven 1959), but that rather, particularly in the sciences of human action, the search for explanations of social events may be regarded as concluded when historically and locally effective individual causes (Mayntz 2002) or non-causal "reasons" have been hit upon that could

In memory of Andreas Balog (1946–2008).

have moved actors to initiate definite courses of action (Louch 1966); and, related to this criticism, the most far-reaching objection was that valid explanations do not constitute arguments and do not therefore require a logically deductive connection between the explanans and the explanandum, with the consequence that "pragmatic" (Bromberger 1966), "normative" (Dray 1957), "narrative" (Danto 1965: 233ff.) or "practical" explanations (Wright 1971) have to be admitted. All these criticisms were confronted by orthodox counter-arguments (Hempel 1965: 412ff.), which reinforced the impression that the decades-long debate on the logic of scientific explanation had ended up inconclusively (Koertge 1992). What had gone wrong?

In order to answer this question we should accept two points: as long as we wish to cleave to the possibility of (social scientific) explanations their deductive character should not be impugned, for we in fact "explain" a (social) state of affairs by logically deducing the proposition describing it from an explanans; and additionally, that in order to judge the validity of such deductions we need *laws* indicating which factors ultimately "produce" or "generate" a relevant event (Bunge 1979).[1] If, as not only I assume (Hedström 2005: 67), the task of social scientific analyses consists in the explanation of macro-structural distributions, collective effects of action, forms of organization and relationship, the functioning of systems of action, or, in short, "collective phenomena" (Popper 1966: 98), then we must pose the question about what the laws are with whose help we can deduce such macroscopic explananda.

The possible answers to this question can, I think, be classified in the following manner. Some theorists seek the "social laws" governing the (historical) course of macroscopic processes and firmly adhere to the idea that an explanatory social science could arise only if it succeeded in finding the "developmental laws," "laws of motion" or of "transformation" and "transition" of society, or at least in identifying the conditions of social-structural equilibria. Against this spoke the fact that no one has ever been able to discover such structural and process laws (and this for purely logical reasons (Popper 1961: v)), or respectively that all candidates have continually been empirically proven to be false (Schmid 2006: 12f., 139ff.). If theorists did not prefer to ignore these results so as to continue to search for such laws (McIntyre 1996), two responses were possible. The first was that they abandoned the search for social laws and restricted themselves to laws of individual action; the best-known version

[1] Whether causal processes can be (successfully?) manipulated by intentional actors (Woodward 2003) is of quite secondary importance for an appropriate understanding of the conception of "causality."

of this approach was represented by so-called "behavioral sociology," according to which macroscopic phenomena had to be "reduced" to individual actions (Homans 1974), which was naturally bound up with the denial of the action-directing influence of structures and left unilluminated how such structures could arise from the actions of individual actors. The legitimate concern of traditional macro-sociology, to explain the "behavior" of social systems (Coleman 1990), was thus increasingly lost sight of. The second response was to renounce the search for such laws of individual action, and so for all nomological connections of any description, with the immediate consequence that the social sciences have no explanatory task at all, but rather can and must devote themselves to an (non-explanatory) method of "understanding," or confine themselves to the construction of descriptive concepts or to typologies of social processes produced with their help.

It is my belief that we can ward off this looming self-dissolution of the social scientific program of explanation if we follow a mediation proposal which may be traced to Robert King Merton (and his students) (Schmid 1998: 71ff., Schmid 2006: 53ff.) as far as sociology in particular is concerned, but which has also met with increasing acceptance in neighboring disciplines (Mahoney 2001). I shall confine myself to a systematic reconstruction of this proposal, which I should like to discuss under the heading of "microfoundational explanations in the social sciences."

The logic of microfoundational explanations in the social sciences

In my view, every microfoundational explanatory practice is marked by the following confrontation: on the one hand, the belief in the existence of an all-comprehending social theory, which could state all the laws of social processes on the basis of an imperial system of concepts, cannot be realized if only for the reason that there are no such laws; on the other hand, recourse to a reductively applied theory of behavior is insufficient to explain the emergence and independence of social phenomena, because the laws of such a theory contain no structural predicates and therefore say nothing about "social relations." In order to explain such "phenomena" we need rather, as Robert K. Merton has expressed it, an "analysis" of the "social interrelations (of persons)" (Merton 1964: 56) in which the functioning and consequences of such relations should be grasped as the often unintended collective product of the actions of goal-oriented and intention-guided actors who must choose among structurally pre-formed alternative actions (Stinchcombe 1975).

Such an explanatory guideline can be fruitful if the following assumptions can be satisfied: first, it must be established that the individual action of each actor can be explained as the consequence of a *choice*, which implies that he knows how to reach a decision about his action in view of his established goals and subjective information. An approach to such a theory of individual choice is possible along various paths. The theory defended by most authors specifies as the most important elements of choice the possibility and capacity of an actor to use his resources intelligently and creatively, to expect and evaluate target states, and thereby to be able to resort to a decision-algorithm which identifies at least one of the envisaged alternatives as the best possible or most cost-efficient result of his considerations. Thus the core of such a theory is an actor who has (decision relevant) "capacities" and knows how to apply them to the planning and execution of his action in view of his (perceived and differently evaluated) possibilities and restrictions. The RREEMM theory of decision (Lindenberg 1985), the value-expectation theory (Esser 1993), the use of rational choice theory (Becker 1976), various theories of "boundedly" rational action (Simon 1983), the theory of cognitive dissonance (Kuran 1998) and learning (Homans 1974), the prospect theory (Kahneman and Tversky 1979), the "theory of frame-selection" (Esser 2003), the theory of "good reasons" (Boudon 2003), of "folk psychology" (Balog 1989) or of action guided by "desires and beliefs" (Hedström 2005), but also psychoanalysis (Alexander 1968), put forth comparable and (as I think) logically compatible explanatory proposals (Balog 2001: 365).

As has been acknowledged, however, the object of social scientific explanations is not the individual actions of individual actors, but "social states of affairs" or "collective phenomena." In my view, the discussion of the logical character and substantive goals of social scientific explanations has hitherto been afflicted by the ambivalent use of these concepts in various theories. By a "collective fact," interactively minded theorists mean first of all the fact that – as a consequence of a division of labor – actors must enter into *interdependent relationships* which can take at least two forms (Boudon 1979). In the first case, the actors restrict themselves to observing the action of others and to adjusting their actions adaptively to the former; it is here (at least implicitly) supposed that they possess the undisputed *right* to act in a self-interested manner. In the second case, the actors are not accorded this right and they are therefore constrained to take the behavioral challenges of others into account, especially if they must reckon with interventions, should they disregard the interests of their co-actors. The implication of both forms of "action-orientation" is that actors can construct and direct

their relationships by means of the reciprocal recognition of rights (and logically bound up with this, of norms) (Coleman 1990), so that we may infer that "social phenomenon" means in the first place "regulated forms of relationship." *On the other hand*, the collective consequences or, as they are called, the "structural effects" (Blau 1977: 144ff.) or "composition effects" (Boudon 1977: 271; Boudon 1986: 56ff.) of their interdependent (as well as regulated) action and their (sometimes called "emergent" (Sawyer 2005)) distributional characteristics are also regarded as "social facts." It is "social facts" of *this kind* that particularly structuralist theorists have in mind when they direct their attention partly to how these effects can gain in independence against individual actors (Blau 1994) and partly to how they emerge from the individual action of a plurality of intentional and self-interested actors (Wippler 1978). In recognition of the fact that *both kinds* of "social phenomena," *interdependences* and *distributional effects*, play a determinative role in action, we must (obviously) take care that a serviceable social scientific explanatory model makes both kinds intelligible as a (possibly unintended, sometimes undesired and unexpected) collective consequence of intention-guided individual actions.

I think that the previous reflections suffice to distill the *logic of a social scientific explanatory argument*. It will have become clear that we cannot understand such explanations as one-level subsumptions which would allow derivation of the consequences of individual rule-oriented action *directly* from a theory of action (of whatever sort). Instead, we should assume that there are obviously *four separate steps of explanation* to be distinguished. The *first step* is to explain the *actions of individual actors* with recourse to the internal, acquired and genetically determined capacities by means of which they perceive and evaluate their situation, while the success of any action depends on the possibilities and restrictions confronting each individual actor. In order to identify these opportunities we must take into account that there are two aspects to the conditions of success of any action. On the one hand, in order to organize and project their action, actors must be able to draw on ("material") resources which (in many cases) they can regard as unquestioned "data" in weighing their decision problem. Any certainty is instantly lost when they are forced to heed whether and to what degree their in principle unpredictable co-actors can co-condition, if not disrupt or hinder, their plans. Or differently stated, the action of others is one of *the* major opportunities of the action of each actor, which he must include in his individual decisions. Every "social action" that is realized in this manner may be understood (as game theory proposes) as a "strategic action," whereby (in the present case) we leave open

whether the actors communicate and influence each other through direct, expectation-guided interaction or by means of the "externalities" they might have produced.

This recognition of the interdependence of their action is important because the admissibility and inevitability of the *second step of explanation* depends upon it. This further step consists in determining how the various individual actors link their action with each other in such a way that reproducible action-constellations can emerge and prevail. Following a widely discussed proposal, we may call these processes of linkage a "social mechanism" (Lindenberg 1977), which may be analyzed under certain presuppositions. First, we must see that there are different such mechanisms and, second, that the demand for them and their chance of success depend upon the *extent* and *nature of the problem* with which the actors are confronted in their attempt to assert or to adjust their mutual interests in strategic situations (cf. Ullmann-Margalit 1977 who differentiates between problems of "coordination," "cooperation" and "partiality"). In furnishing the required definition of what may be regarded as such a "problem of mutual action adjustment" (Schmid 2006) or a problem of "concerting" interdependent decisions (Ullmann-Margalit 1977: 82), I believe it has been fairly clearly shown that we cannot do without a theory of action by which to explain individual action as an opportunity-conditioned, intention-guided action interested in the best possible returns. Differently stated, it is *only* in light of a theory of action that we can *discover* with what positive or negative collective consequences the actors may reckon when they decide in a specific manner for or against coordination, cooperation or conflict with others. In this connection, many authors believe that game theory is suited (Esser 2000a: 27ff.; Little 1991: 51ff.; Mayntz 2004) to furnish a clear definition of the (logically) expectable constellations of pay-offs with which "rational" or self-interested actors meet in strategically interdependent situations. The more detailed situation-logical analysis[2] of such reciprocally interlocking options of action should be prepared to meet repeatedly with constellations in which the observed actors have to accept losses owing to the antagonistic irreconcilability of their objectives and expectations, partly because their susceptibility to opportunism and deceit prevents the attainment of an optimal securing and distribution of common gains, and partly because (sometimes intolerable) set-up

[2] Popper's "logic of the situation" is insufficiently developed insofar as he deals only with "games against nature" and not with "mechanisms of mutual action accommodation" (Hedström *et al.* 1998).

and transaction costs accrue even when the actors have their eyes on yields that are as mutually beneficial as they are undisputed.

The precise determination of the cost structure with which actors see themselves confronted is, however, only one necessary condition for their desiring to become engaged in establishing and maintaining mechanisms of a certain form; equally important is whether they can activate motivational reasons that incline them to come to terms with their co-actors. Here, in addition to relatively unchanging basic demands, several influential factors play a role. Above all it will be important what kind of return the actors may reckon with when they decide to pay attention to the interests of others. With this, in turn, we broach the multi-level subsequent question of whether the actors can expect private or collective goods as a result of their efforts at reciprocal adjustment of actions, whether these goods can be shared without difficulty or are subject to competitive consumption, to what extent the quality of the goods can be checked or foreseen, whether the rights of use of such goods can be transferred wholly or only in part, and other questions. In addition, we must know the initial allocation of rights or "possessions" over which the actors dispose. In view of prevailing inequalities of power and divergences of distributive interest, it would be naive to assume that all participating actors are satisfied with every allocation; that is to say, every modeling of such relations should take into account possible sub-optimal results of distribution and their corresponding criticism and susceptibility to revision.

If these presuppositions are clarified, then the social analyst may hope to discover whether and with what prospects of success a specific regulation-relevant mechanism can prevail (Schmid 2004: 247ff.); for instance, whether the establishment of *exchange relationships* is suitable for mutually concerting their interest in returns, or whether rights of *authority* should be granted in order to effect the compatibility of action, or whether an actor is better off when he aims at a *moral self-obligation*. In all cases, we must ascertain whether collective decisions are required in order to effect the corresponding problem solutions or whether the private determination of courses of action is allowed (Coleman 1986: 15ff.), how far the use of influence and violence yields returns (Boehm 1987; Gambetta 1993), whether contracts must be concluded (Schweitzer 1999), whether trust (Hardin 2002) or some kind of "social capital" can be accumulated (Bourdieu 1992: 46ff.), whether damages can be compensated (Sened 1997), and other issues. The evident diversity and openness of this catalog of conditions suggests the thesis that, as a rule, it is the theoretically quite non-transparent interlocking of *various* procedures of co-adjustment that admits of the hope that actors, in spite of their inalterable ignorance, ill-mindedness

and unscrupulousness can find a satisfactory safeguard for their need of co-adjustment at least for a time.

Whether such a permanent stabilizing of mutually adjusted interaction can be enforced, however, is an open question. In fact, from its beginnings, social theory has been occupied with the evidently not conclusively soluble problem under what circumstances actors can succeed in maintaining entrenched (or, as they are called, "functional" or "organic") forms of relationships, or whether they see themselves compelled to refashion or even to abandon them. What is certain, of course, is that in order to answer the question about the "social order" two *further steps of explanation* are needed. First, we must consider what has recently been discussed under the name of the "aggregation problem." The solution of this problem, which has to be given for every possible theory of action trying to explain collective action results (Balog 2001: 169ff.), requires developing an idea of how to identify the "collective consequences" of common co-adjustment attempts that actors have to reckon with. The exact logical character of such aggregations is still under dispute (Schmid 2009a). But it seems to be clear that mechanismicly regulated actions are not the mere result of formal-analytical "transformation rules" (Esser 2000; Lindenberg 1977), but rather (non-analytical) causal consequences of (possibly quite defective, even anomic) legal or normative regulations that underlie a certain mechanism. That is to say, the rule-based operation of an entrenched mechanism must already be known before its aggregate consequences can be identified.

The identification of the consequences of collective actions, however, does not determine how actors should conduct themselves toward these consequences and whether they are at all able to react to them. In order to complete the task of explanation, therefore, we need a *further* (and final) *step*. In order to judge how aggregate or "compositional effects" of their collective action affect their subsequent decisions, and so the probable reproduction or reorganization of an investigated mechanism, the researcher needs additional information about their "*recursive*" *effects* (Luhmann 1997), which, precisely because actors are often insufficiently informed particularly about the hidden effects of their inter-related actions, can be acquired only when he knows how to discover the direction in which the collective effects of such mechanisms affect the subjective willingness as the possibilities of the actors to continue to align their action.

Interpretations

The analysis of mechanismic explanatory arguments into four steps requires several comments: some concerning the compatibility with

Hempel's original explanatory model, and others that are intended to throw light on the heuristics upon which proponents of such an understanding of explanation believe they can draw.

Multi-level analysis and deduction

To begin with, it should now be evident that the explanatory pattern in question diverges from the simple Hempel model inasmuch as it insists that social scientific explanations constitute *tiered multi-level explanations* (Hedström 2005: 35). They correspond to Hempel's explanatory logic only in the *first step of explanation*, where it is a question of explaining the actions of individual actors. To this end, two convergent requirements need to be satisfied: *on the one hand*, every explanation of individual action should be able to appeal to *nomological assumptions* about *how* individual actors *determine* a certain action. A strictly individualistic theory of action, which possesses a causal character so far as it designates certain processes that "energize" individual action and specifies the capacities, points of view and evaluations of problems and possibilities by means of which actors reach decisions, describes the relevant (psychological) mechanisms, whose focused investigation has been recommended above all by Jon Elster (Elster 1979, 2000). As far as can be seen, all the theories of action that come into question treat the choices of actors as their own (purely) individual, intentional and self-directed internal activity, which must be explained *as such*. I assume that our nomological knowledge (in the strict sense relevant to the social sciences) is related exclusively to the genesis or "selection" (Esser 1993: 120f.) of individual actions, and not, for instance, to their resultant "social relations" and their effects and repercussions.

This thesis has, as I believe, a non-trivial consequence: if it is true that we have no knowledge of social laws that would allow us to deduce a social-structural explanandum directly from an explanans, then it is not possible to construct social scientific explanations without *recourse* to an underlying theory of action; social scientific explanations ("social phenomena") therefore *necessarily* refer to an action-theoretical and nomological "core" (Esser 1993: 95; Esser 2004: 34, 37). Or in another formulation, social phenomena may be regarded as having been explained *only* when their genesis, operation and reorganization (in the last instance) is accounted for on the basis of the individual adaptive actions of individual actors (Lindenberg 1977, 1992; Little 1998), and this only when the conditions of the eventual success of their collective action are not "defined into them," as sociologists are wont to do (Campbell 1996). Corresponding to the basic methodological ideas

of individualism, this demands that social or macroscopic explananda located at a certain level n (or $n + x$) must be explained in the light of a (causal) theory of individual choice operating at the level $n - 1$; in this way, social scientific explanation operates *microfoundationally* or, as it is called in other places, in the form of "depth explanations" (Bunge 1967: 26ff.).

This is to say that the formulation of such depth explanations is possible only when (as suggested) a *second condition* is satisfied. Since theories of action designate exclusively the psychological mechanisms of action selection, for the completion of social scientific explanations we need, in addition to the antecedent conditions of the (individual) theory, information about the actors' action-situation and the resultant problems. Such information must be introduced in the form of (situation-specific) *additional hypotheses* that recognize the (independent, supra-individual) influential qualities of structural conditions and forego their *logical* reduction to assumptions about action or behavior. Instead of performing such a reduction, we must assume that the actors' external or situational circumstances "canalize" their decision acts in a contingent manner, and not in a direct (mentally unmediated) "causally effective" sense, as we have assumed for the (psychologically directed) genesis of their action. The first step of a multi-level explanation succeeds, therefore, only when we can refer to testable hypotheses about the situational (or "structural") opportunities of actors, independently of the given theory of action.

With the introduction of such situational assumptions, we enter, in terms of theoretical technique, the (first) level of (theory guided) *modeling*. That is, in applying a theory of action to the actual decision-situations of actors we formulate a "situation model" of (in principle and unavoidably) contingent and highly various conditions and contexts that guide individual action, whereby we continue to assume that this modeling will not bring to light any (deterministic or inductively won) laws pertaining to the peculiarities of social situations. If moreover we are interested in answering the question of how these (generally problematic) contexts arise from the action of *a plurality* of actors whose action is to be explained individually, we must have recourse to *supplementary hypotheses* about how and under what conditions a disruption-free coadjustment of their action comes about. Among such hypotheses is not only the assumption that recourse to reliable rules will be sensible to this end, but also an idea of under what circumstances self-interested actors are prepared to show the willingness to establish and adhere to such rules, or (if required and possible) to revise them (Baurmann 1996; Schmid 1998: 118ff., 131ff.). This means, however, that here at least we

should have to designate an interdependence-directing mechanism or a rule-based institution (in the sense of Douglass North 1990), that is an action regulating device that we should have to grasp as a highly contingent and accordingly variable process of inter-personal adjustment which, operating as a (restrictive) "social fact," furnishes the possible freedom of action of individual actors with absolute limits. Thus the explanatory task consists in developing a "process model of a mechanism" that designates the social (or "structural") framework within which self-interested actors are reciprocally forced to renounce seeking their maximum advantage where this tends to be bound up with ("morally" unacceptable (Bayertz 2004; Wilson 2004)) damage to the interests of their co-actors.

The *third* and *fourth steps of explanation* then focus on the fact that the existence of certain mechanisms entails non-reductive (or "emergently" operating) collective (or aggregate) effects which, mediated by their retroactive influence on the possibilities of action, situational perceptions and motives of actors (Hedström 2005: 42ff.), affect their further decisions and, in consequence, the probabilities of the persistence or transformation of the relevant mechanisms. Important is that, owing to the indefiniteness of individual projects of actions, we also do not meet here with "laws" of a generalizable kind; instead, in order to identify collective action consequences and their recursive effects, we need *contingent hypotheses* that supply us with information which must (and can) extend beyond what we have to know about the establishment and operation of the mechanisms themselves.

The logical form of the explanatory argument to be constructed in accordance with these considerations is no longer merely multi-level, but also (as has already been suggested and as Hempel's model demands) *deductive* (Hedström 2005: 30f.). This holds for both the explanation of individual actions and for the subsequent steps in which we, proceeding from assumptions about action and situation with respect to a plurality of actors, seek (logically necessary) deductions about possible interdependences of the former, or about the resultant consequences for the formation of mechanisms, and finally about their retroactive collective effects. By building these four steps of explanation on each other, we can perfect and at the same time complete the explanatory argumentation, whereby the traversal of the entire argument is logically bound up with the fact that the factors which have been treated in each previous step can function as parameters or boundary conditions for the subsequent steps. It should be noted that (as has been frequently suggested) each further explanatory step requires *auxiliary assumptions* about the peculiarity of the circumstances specific to each level, assumptions whose

selection is in turn possible only when we have already determined the explanandum of the next given step and which (fortunately) compels us to specify our (model relevant) explanatory intentions and problems.

This (logically complete) linkage of the various steps of explanation has two implications. The first is that, in this way, static analyses of the operation and collective consequences of action-canalizing mechanism can be converted into *dynamic models* describing the often indeterminate behavior (Hardin 2003) over time of corresponding systems of interdependence, which allows us to treat "structural reproduction" and "structural change" within the framework of one and the same set of theoretical guidelines (as Boudon and Luhmann called for long ago). The second implication is that, as long as the (mainly unintended and indefinite) consequences of their efforts find a solution of their action problems repeatedly offer them new possibilities of and stimuli to changed (or even innovative) reactions, conclusive or ultra-stable equilibria of their forms of social intercourse will be improbable (or even impossible). In this way, we can leave behind us the constrictions of the traditional functionalist "theory of social order" (in sense of Durkheim and Weber and, above all, of Parsons and his "school").

On the heuristics of mechanismic explanation

From the logic of a microfoundational explanatory argument of the kind here described and the related technique of argumentation and modeling, I believe we can draw several *heuristic rules* for how to proceed if we are interested in developing a theory-guided and at the same time testable program of sociological research.

First of all, no one should feel obliged to carry out all the steps of explanation at the same time or to strive to treat *simultaneously* all the variable factors that can be taken into account at each level of modeling. It is important to specify only which influencing factors we posit as constant, equal to 1 or place as a model parameter before the brackets, because otherwise we can neither judge the argumentative fidelity of an explanatory step nor ensure its testability. That is, I defend as Imre Lakatos (1970: 101f.) or Nancy Cartwright (1989: 161ff.) did, the controlled use of so-called "ceteris paribus clauses."

As we must introduce additional hypotheses in every step of explanation in order to proceed with our deduction, there is reason to consider *whence* we can draw these additional theses. In many cases we have to invent them anew and the fruitfulness of a research program may depend on our capacity to do so; but if we can borrow them from other models, it is then possible to connect (logically) the work at hand on

our own sub-model with corresponding parallel attempts. In this way our modeling acquires the open, "set-theoretical" character described by Heelan (1981). The further question arises of whether we must test these additional assumptions ourselves. A conclusive answer will depend upon what we wish to know; sometimes it may suffice to clarify whether a desired explanandum is deducible only if the relevant additional hypotheses *would* be true, or conversely the failure of a possible prognosis can be explained by its failure (contrary to our covert expectations) to fulfill the relevant conditions. The task is then to examine these conjectures more closely, and this leads the continuing research in an assignable direction, while at the same time fulfilling the conditions of Lakatos' "positive heuristics" (Lakatos 1970: 135).

Moreover, every sub-model allows deductions from hitherto unconsidered implications, which can be tested in order to examine the explanatory and truth-value of our premises. Here I count on the possibility that the attention of empirical research can be rerouted (as in the case of the additional hypotheses discussed in the previous section) from the usual inductive (and correspondingly non-theoretical) collection of data to a selection of topics relevant to our modeling (Esser 2004: 28ff.; Hedström 2005: 114ff.). This would open a theory-guided field of empirical research that satisfies the usual standards of a realist, truth-guided and insofar critical methodology.

The requirement of examination also applies to the theory of action itself. A special difficulty results in this connection from the fact that the social sciences have up to now been unable to agree on what assumptions about action they wish to take as the basis of their explanations and how the various proposals (logically) fit together. I believe, with Esser (2003: 70f.), that all previous considerations about a ("general") theory of human action can be "synthesized" (Schmid 2004: 24ff.; Schmid 2006); but whether we should endeavor to achieve such an "integration" of various "paradigms" depends in equal parts on whether we can have recourse to an effective comparative methodology (Schmid 2009) and how far the extension (or "complication") of our premises about action may yield *additional* insights at the level of structural connections in our attempts to found mechanismic and structural links in a theory of action (Lindenberg 1992: 19; Stinchcombe 1993: 35). If such additions are possible and desirable, then it makes no methodological sense to exaggerate the protection of a "hard core" of action theory and to forego its conceivable "extensions" (Stegmüller 1980: 377). At the same time it is perfectly legitimate to leave the set of action assumptions unchanged in order to pursue an identifiable "research program" (in the sense of Lakatos 1970).

Last but not least, the heuristics of a mechanismic explanatory program set us of course at liberty to investigate those mechanisms that, for whatever reasons, happen to interest us, and we cannot exclude that it may be wise to mark off research programs from each other by asking each to dedicate itself primarily to one of the various mechanisms. The unhesitating adherence to this rule explains the relatively uncontroversial (if now only vestigially recognizable) division of the social sciences into special disciplines. Such independently operating disciplines should, however, be wary of maintaining that all problems of interaction can be solved by means of a single mechanism. This claim not only sounds "imperialistic"; it is, in view of the existence of very different procedures of solving action problems, evidently false. On the other hand, since the investigation of the co-action of various mechanisms, even if possible, constitutes a heavy burden, it would probably be reasonable to concentrate various research programs on the supervised elaboration of generalizable "structure models" (Esser 2002) of individual action mechanisms – a project that we may hope to promote without each individual (and highly circumscribed) research interest seeking to profile itself by belittling neighboring programs as naive, irrevocably false or irrelevant.

REFERENCES

Alexander, Peter. 1968. "Rational behaviour and psychoanalytical explanation," in Norman S. Care and Charles Landesman (eds.), *Readings in the Theory of Action*. Bloomington and London: Indiana University Press, 159–78.

Balog, Andreas. 1989. *Rekonstruktion von Handlungen*. Opladen: Westdeutscher Verlag.

2001. *Neue Entwicklungen in der soziologischen Theorie*. Stuttgart: Lucius & Lucius.

2006. *Soziale Phänomene. Identität, Aufbau und Erklärung*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Baurmann, Michael. 1996. *Der Markt der Tugend. Recht und Moral in der liberalen Gesellschaft. Eine soziologische Untersuchung*. Tübingen: J.C.B. Mohr (Paul Siebeck).

Bayertz, Kurt. 1980. *Wissenschaft als historischer Prozess. Die antipositivistische Wende in der Wissenschaftstheorie*. München: Wilhelm Fink Verlag.

2004. *Warum überhaupt moralisch sein?* München: C.H. Beck Verlag.

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

Blau, Peter M. 1977. *Inequality and Heterogeneity. A Primitive Theory of Social Structure*. New York and London: The Free Press and Collier Macmillan.

1994. *Structural Contexts of Opportunities*. Chicago and London: University of Chicago Press

Boehm, Christopher. 1987. *Blood Revenge. The Enactment and Management of Conflict in Montenegro and Other Tribal Societies*. Philadelphia: University of Pennsylvania Press.

Boudon, Raymond. 1977. "Soziale Bedingtheit und Freiheit des Individuums. Das Problem des homo sociologicus," in Klaus Eichner and Werner Habermehl (eds.), *Probleme der Erklärung sozialen Verhaltens*. Meisenheim: Verlag Anton Hain, 214–76.

1979. *Widersprüche sozialen Handelns*. Neuwied and Darmstadt: Luchterhand Verlag.

1986. *Theories of Social Change. A Critical Appraisal*. Cambridge: Polity Press.

2003. *Raison, bonnes raisons*. Paris: Presses universitaires de France.

Bourdieu, Pierre. 1992. *Die verborgenen Mechanismen der Macht, Schriften zu Politik und Kultur 1*. Hamburg: VSA-Verlag.

Bromberger, Sylvain. 1966. "Why-questions," in Robert G. Colodny (ed.), *Mind and Cosmos. Essays in Contemporary Science and Philosophy*. University of Pittsburgh Press, 86–111.

Bunge, Mario. 1967. *Scientific Research II. The Search for Truth*, 3rd rev. edn. Berlin, Heidelberg and New York: Springer.

1979. *Causality and Modern Science*. New York: Dover Publications.

2004. "How does it work? The search for explanatory mechanisms," *Philosophy of the Social Sciences* 34: 182–210.

Campbell, Colin. 1996. *The Myth of Social Action*. Cambridge University Press.

Cartwright, Nancy. 1989. *Natures Capacities and their Measurement*. Oxford: Clarendon Press.

Coleman, James S. 1986. *Individual Interests and Collective Action*. Cambridge University Press.

1990. *Foundations of Social Theory*. Cambridge, MA and London: The Belknap Press.

Danto, C. Arthur. 1965. *Analytical Philosophy of History*. Cambridge University Press.

Dray, William. 1957. *Laws and Explanation in History*. Oxford: Clarendon Press.

Elster, Jon. 1979. *Ulysses and the Sirens. Studies in Rationality and Irrationality*. Cambridge and Paris: Cambridge University Press and Édition de la Maison des Sciences de L'Homme.

2000. *Ulysses Unbound. Studies in Rationality, Precommitment, and Constraints*. Cambridge University Press.

Esser, Hartmut. 1993. *Soziologie. Allgemeine Grundlagen*. Frankfurt and New York: Campus Verlag.

2000. *Soziologie. Spezielle Grundlagen, Band 2: Die Konstruktion der Gesellschaft*. Frankfurt and New York: Campus Verlag.

2000a. *Soziologie. Spezielle Grundlagen, Band 3: Soziales Handeln*. Frankfurt and New York: Campus Verlag.

2002. "Was könnte man (heute) unter einer 'Theorie mittlerer Reichweite' verstehen?" in Renate Mayntz (ed.), *Akteure – Mechanismen – Modelle. Zur Theoriefähigkeit makro-sozialer Analysen*. New York and Frankfurt: Campus Verlag, 128–50.

2003. "Die Rationalität der Werte. Die Typen des Handelns und das Modell der soziologischen Erklärung," in Gert Albert, Agathe Bienfait, Steffen Sigmund and Claus Wendt (eds.), *Das Max Weber-Paradigma*. Tübingen: Mohr Siebeck, 153–87.

2004. *Soziologische Anstöße*. Frankfurt and New York: Campus Verlag.

Gambetta, Diego. 1993. *The Sicilian Mafia. The Business of Private Protection*. Cambridge, MA and London: Harvard University Press.

Hardin, Russell. 2002. *Trust and Trustworthiness*. New York: Russell Sage Foundation.

2003. *Indeterminacy and Society*. Princeton and Oxford: Princeton University Press.

Hedström, Peter. 2005. *Dissecting the Social. On the Principles of Analytical Sociology*. Cambridge University Press

Hedström, Peter, Richard Swedberg and Lars Udéhn. 1998. "Popper's situational analysis in contemporary sociology," *Philosophy of the Social Sciences* 28: 339–64.

Heelan, Paul A. 1981. "Verbandstheoretische Betrachtung des Erkenntnisfortschritts," in Gerard Radnitzky and Gunnar Andersson (eds.), *Voraussetzungen und Grenzen der Wissenschaft*. Tübingen: J.C.B. Mohr (Paul Siebeck), 339–46.

Hempel, Carl G. 1965. *Aspects of Scientific Explanation and other Essays in the Philosophy of Science*. New York and London: The Free Press.

Homans, George C. 1974. *Human Behavior. Its Elementary Forms*, 2nd edn. New York: Harcourt Brace Jovanovich.

Kahneman, Daniel and Amos Tversky. 1979. "Prospect theory. An analysis of decisions under risk," *Econometrica* 47: 263–91.

Koertge, Noretta. 1992. "Explanation and its problems," *The British Journal of Philosophy of Science* 43: 85–98.

Kuran, Timur. 1998. "Social mechanisms of dissonance reduction," in Peter Hedström and Richard Swedberg (eds.), *Social Mechanisms. An Analytical Approach to Social Theory*. Cambridge University Press, 147–71.

Lakatos, Imre. 1970. "Falsification and the methodology of scientific research programmes," in Imre Lakatos and Alan Musgrave (eds.), *Criticism and the Growth of Knowledge*. Cambridge University Press, 91–195.

Lindenberg, Siegwart. 1977. "Individuelle Effekte, kollektive Phänomene und das Problem der Transformation," in Kurt Eichner and Werner Habermehl (eds.), *Problem der Erklärung sozialen Verhaltens*. Meisenheim: Verlag Anton Hain, 46–84.

1985. "Rational choice and sociological theory. New pressures on economics and social science," *Zeitschrift für die gesamte Staatswissenschaft* 141: 244–55.

1992. "The method of decreasing abstraction," in James S. Coleman and Thomas J. Fararo (eds.), *Rational Choice Theory. Advocacy and Critique*. Newbury Park, London and New Delhi: Sage Publications, 3–20.

Little, Daniel. 1991. *Varieties of Social Explanation. An Introduction to the Philosophy of Social Science*. Boulder: Westview Press.

1998. *Microfoundation, Method, and Causation*. New Brunswick and London: Transaction Publishers.

Louch, A.R. 1966. *Explanation and Human Action*. Oxford: Basil Blackwell.
Luhmann, Niklas. 1997. *Die Gesellschaft der Gesellschaft*. Frankfurt: Suhrkamp Verlag.
McIntyre, Lee C. 1996. *Laws and Explanations in the Social Sciences. Defending a Science of Human Behavior*. Boulder: Westview Press.
Mahoney, James. 2001. "Beyond correlation analysis. Recent innovations in theory and method," *Sociological Forum* 16: 575–93.
Manicas, Peter T. 2006. *A Realist Philosophy of Social Science. Explanation and Understanding*. Cambridge University Press.
Mayntz, Renate. 2002. "Zur Theoriefähigkeit makro-sozialer Analysen," in: Renate Mayntz (ed.), *Akteure – Mechanismen – Modelle. Zur Theoriefähigkeit makrosozialer Analysen*. Frankfurt and New York: Campus Verlag, 7–43.
2004. "Mechanisms in the analysis of social macro-phenomena," *Philosophy of the Social Sciences* 34: 237–59.
Merton, Robert K. 1964. *Social Theory and Social Structure*. New York: The Free Press.
North, Douglass C. 1990. *Institutions, Institutional Change and Economic Performance*. Cambridge University Press.
Popper, Karl R. 1961. *The Poverty of Historicism*. London: Routledge and Kegan Paul.
1966. *The Open Society and its Enemies, Vol. II. The High Tide of Prophecy: Hegel and Marx*, 5th edn. London: Routledge & Kegan Paul.
Sawyer, Keith. 2005. *Social Emergence. Societies as Complex Systems*. Cambridge University Press
Schmid, Michael. 1998. *Soziales Handeln und strukturelle Selektion. Beiträge zur Theorie sozialer Systeme*. Opladen: Westdeutscher Verlag.
2004. *Rationales Handeln und soziale Prozesse. Beiträge zur soziologischen Theoriebildung*. Wiesbaden: VS Verlag für Sozialwissenschaften.
2006. *Die Logik mechanismischer Erklärungen*. Wiesbaden: VS Verlag für Sozialwissenschaften.
2009. "Theorien, Modelle und Erklärungen. Einige Grundprobleme des soziologischen Theorienvergleichs," in Gerhard Preyer (ed.), *Neuer Mensch und kollektive Identität in der Kommunikationsgesellschaft. Karl Otto Hondrich zum Gedächtnis*. Wiesbaden: VS Verlag für Sozialwissenschaften, 323–59.
2009a. "Das Aggregationsproblem. Versuch einer methodologischen Analyse," in Paul Hill, Frank Kalter, Johannes Kopp, Clemens Kroneberg and Rainer Schnell (eds.), *Hartmut Essers Erklärende Soziologie, Kontroversen und Perspektiven*. Frankfurt and New York: Campus Verlag, 135–66.
Schweitzer, Urs. 1999. *Vertragstheorie*. Tübingen: Mohr Siebeck.
Scriven, Michael. 1959. "Truisms as the grounds for historical explanations," in Patrick Gardiner (ed.), *Theories of History*. New York and London: The Free Press and Collier Macmillan, 443–75.
Sened, Itai. 1997. *The Political Institution of Private Property*. Cambridge University Press.
Simon, Herbert A. 1983. *Reason in Human Affairs*. Palo Alto: Stanford University Press.

Stegmüller, Wolfgang. 1980. *Neue Wege der Wissenschaftsphilosophie*. Berlin, Heidelberg and New York: Springer Verlag.

Stinchcombe, Arthur L. 1975. "Merton's theory of social structure," in Lewis A. Coser (ed.), *The Idea of Social Structure. Papers in Honor of Robert K. Merton*. New York, Chicago, San Francisco and Atlanta: Harcourt Brace Jovanovich, 11–33.

1993. "The conditions of fruitfulness of theorizing about mechanisms in the social sciences," in Aage B. Sørensen and Seymour Spilerman (eds.), *Social Theory and Social Policy. Essays in Honor of J.S. Coleman*. Westport and London: Praeger, 23–41.

Ullmann-Margalit, Edna. 1977. *The Emergence of Norms*. Oxford: Clarendon Press.

Wilson, Catherine. 2004. *Moral Animals. Ideals and Constraints in Moral Theory*. Oxford: Clarendon Press.

Wippler, Reinhard. 1978. "The structural-individualistic approach in Dutch sociology. Towards an explanatory social science," *The Netherland Journal of Sociology* 14, 135–55.

Woodward, James. 2003. *Making Things Happen. A Theory of Causal Explanation*. Oxford University Press.

Wright, Georg H. von. 1971. *Explanation and Understanding*. London: Routledge and Kegan Paul.

8 Social mechanisms and explanatory relevance

Petri Ylikoski

In this chapter I will discuss mechanistic explanation in the social sciences from the point of view of the philosophical theory of explanation. My aim is to show that the current accounts of mechanistic explanation do not serve the agenda of analytical social science as well as they should. I will not challenge the idea that causal explanations in the social sciences involve mechanisms or that social scientists should seek causal explanations and a mechanistic understanding of social phenomena, but will argue that to improve explanatory practices in the social sciences analytical social scientists should employ tools more substantial than the metaphor of a mechanism.

I will begin by presenting the basics of the erotetic approach to explanation and the notion of explanatory understanding. The second section of the chapter argues that mechanisms do have many important roles related to explanation, but that they do not provide a solution to the problem of explanatory relevance. The third section introduces the idea of a contrastive *explanandum* and argues that paying attention to the identity of the *explanandum* is a necessary condition for proper assessment of explanatory claims. The fourth section argues that explanatory relevance should be understood as counterfactual dependence and that a manipulationist account of causation provides a fruitful framework for making sense of causal explanations. The fifth section discusses some of the consequences of these ideas for mechanistic explanation in the social sciences.

The purpose of the theory of explanation

The aim of the theory of explanation is to make sense of explanations. It addresses questions such as following: what kind of an activity is explanation and how is it related to other epistemic (and practical) activities? What kinds of explanations are there and what are their relationships to each other? By which criteria are explanations evaluated by scientists and by which criteria should they be evaluated? All of these questions

circle around the central question: *what constitutes an explanation?* A convenient way to approach this question is to think about *explanatory relevance*.

The problem of explanatory relevance is one of the major challenges for the covering-law account of explanation. The examples of men taking contraceptive pills explaining them not becoming pregnant, and hexing explaining why salt dissolves in water, are counterexamples of the covering-law account because they involve intuitively explanatorily irrelevant factors while fulfilling all the criteria of a satisfactory covering-law explanation. These counterexamples and other arguments have led philosophers of science to conclude that the covering-law account is a failure. It does not solve the central problem for any theory of explanation: the problem of explanatory relevance. This problem is not easy to solve. For example, Wesley Salmon's causal theory of explanation falls victim to exactly the same counterexamples – it is not enough that we demand that an explanation only provides some information about the causal process; we want to have *relevant* information (Hitchcock 1995).

The problem of explanatory relevance can be understood as a problem of *explanatory selection* (Hesslow 1983). Causal explanation provides information about causal history, but not all information about that history is regarded as explanatory. We have to pick *the right aspects of the causal process* to be included in the explanation. That is to say, how far in the causal history should we reach? How do we choose the events to be included in the explanation? How do we choose the right level of abstraction for describing these events? In how much detail should the events be described and which of their details should be included in the description? Apparently we somehow manage to solve these problems intuitively when we are constructing explanations, but it would be much better if we could make the principles governing these judgments explicit. This is the main challenge for the theory of explanation.

The theory of explanation should not be a purely theoretical enterprise; the ultimate motive is to develop conceptual tools for improving explanatory practices in the sciences (and in everyday life). The theory of explanation should have practical relevance and its success should be judged by the improvements it makes possible in explanatory practices. Such contributions are especially important in the social sciences where controversies about explanation are common. The social mechanisms movement is motivated by similar practical goals. While it has provided philosophical arguments for mechanistic explanation, the aims have been ultimately practical. The improvements in social science explanatory practice are the final evaluation criteria for philosophical arguments about explanation and causation.

In this chapter I will outline an account of explanation that in my judgment best advances the aims of both the philosophical theory of explanation and the social mechanisms movement. I call it *the contrastive counterfactual account of explanation*. The first element of this account is the erotetic approach to explanatory inquiry.

Explanations as answers to questions

Most philosophers of explanation agree that explanations are answers to questions. Some, such as Hempel (1965), have used it as an informal starting point for their discussion, whereas others (Achinstein 1983; van Fraassen 1980) have built their theories of explanation around it. The latter approaches are commonly called *erotetic* approaches to explanation and are often associated with the broader question-theoretical account of scientific research.

In the erotetic approach, scientific enquiry is regarded as a process of answering and elaborating questions. Empirical research is oriented toward fact-finding questions: it aims to answer *what*, *when*, *where* and *how much* questions. The relevance of these questions (and the required precision of the answers) is determined by practical *how to* questions and by explanatory *why* and *how* questions. The insight provided by the erotetic approach is based on the analysis of interrelations between different questions and its description of the research process as an organized series of questions. For example, the research process often requires that big explanatory questions be sliced to a series of smaller ones that can be addressed through empirical inquiry.

Explanation and understanding

In the erotetic approach, explanation is an answer to an explanation-seeking question. The explanation is regarded as complete when it fully answers the given question. This makes the notion of explanation quite narrow. To capture the broader goal of epistemic activities, we need another notion. I suggest that this be *understanding* (Ylikoski 2009). We are interested in finding answers to explanation-seeking questions because we wish to have knowledge about the dependencies governing the world. In other words, the goal of explanation is to understand these dependencies.

Wittgenstein argued that understanding should not be understood as a sensation, an experience, or a state of mind. Understanding is not a special moment or phase, but a more permanent attribute. It is an ability (Wittgenstein 1953: §§ 143–59, 179–84, 321–4). I agree: when a person

understands something, she is able to do certain things, which does not mean that understanding is some sort of special skill. Understanding consists of knowledge about relations of dependence. When one understands something, one can make all kinds of correct inferences about it. Many of these inferences are counterfactual: what would have happened if certain things had been different? What will happen if things were to be changed in a certain manner? Thus the fundamental criterion according to which understanding is attributed is the ability to make inferences to counterfactual situations, the ability to answer contrastive *what-if-things-had-been-different questions* (*what if* questions, for short) by relating possible values of the explanans variables to possible values of the explanandum variable (Ylikoski 2009).

The present account ties together theoretical and practical knowledge: they are not completely different notions. For example, in the case of causal explanation, explanatory understanding is crucial to our pragmatic interests, since answers to *what if* questions concerning the effects of possible interventions enable us to predict the effects of manipulation. Whereas the DN model and the associated epistemic conception of explanation conceive of the possessor of understanding as a passive observer of external events, the contrastive counterfactual theory links our theoretical practices to our roles as active agents (Woodward 2003). The degree of understanding conveyed by an explanation can be defined as the number and importance of counterfactual inferences that the explanatory information makes possible.

Why do many people think that understanding is a mental state or an experience? The reason is that there exists a mental experience that is closely related to understanding: *the sense of understanding*. It is a feeling that tells us when we have understood or grasped something. This sense of confidence (and the feeling that often comes with it) can be easily confused with what we think it indicates: understanding. Ideally these two things would go hand in hand, and assimilating them should not create any trouble. However, real life is different. The sense of understanding is a highly fallible indicator of understanding. Sometimes one has a false sense of understanding and sometimes one understands without having any associated feelings or experiences. The sense of understanding does not give us direct access to knowledge that is the basis of our understanding, so it would be highly surprising if the sense of understanding would turn out to be perfectly calibrated to our understanding. The fallibility of the sense of understanding can be demonstrated experimentally. People often overestimate the detail, coherence and depth of their understanding (Rozenblit and Keil 2002; Ylikoski 2009).

The existence of the sense of understanding should not be regarded as any kind of oddity. It plays an important metacognitive role in our cognitive life. It gives us confidence to try things, and when it is lacking we can sensibly abstain from the activity in question. It also guides the search for new knowledge, and tells us when to stop the search for new information; it signals when we know enough. In addition, the sensation associated with the sense of understanding can have a motivational role. Satisfying curiosity is highly rewarding (Gopnik 2000; Schwitzgebel 1999). It provides motivation for learning and other cognitive activities, and for this reason has an important role in human cognition. The desire to satisfy one's curiosity also provides important psychological motivation for conducting scientific research (Ylikoski 2009).

The phenomenology of the sense of understanding can mislead one into thinking that understanding is an on–off phenomenon ("Now I've got it!"). This is not the case. First, the understanding can be about different aspects of the phenomenon. Second, these aspects may be understood in various degrees. Consider an ordinary object such as a personal computer. Different individuals understand to varying degrees how their computer works. Some might know about the software, or some specific piece of software, and others the hardware. Most people just use the software without any understanding of a computer's internal workings. Despite these differences, they all understand something about their PC. The crucial question is which aspects of it they understand. A comparison of their understanding is possible, but strict assessment of overall understanding is complicated: there are many dimensions to compare. However, the important point is that by asking *what* has been understood, the extent of understanding *can always be specified* (Ylikoski 2009).

This notion of understanding can be used to make sense of social science explanations. For example, we can ask what kind of *what if* questions does a certain account of social mechanism answer? Answering this question can be difficult, but this difficulty only emphasizes the contrast between our confidence that it provides explanatory insight and our inability to articulate (or to agree about) *what* it explains. Surely this kind of exercise would benefit the aims of analytical social science. We can also raise the more general question of whether the social theorist's sense of understanding is well calibrated to his abilities to make correct counterfactual inferences about social phenomena he is studying. I would be highly surprised if it were to turn out that widespread explanatory overconfidence didn't exist among social theorists.

Mechanisms and explanatory relevance

The idea of a mechanism has many uses in the philosophy of science. Most of these ideas also figure in social science discussions about explanation (Hedström and Ylikoski 2010). Here are four different ideas about the contribution of mechanisms to explanatory understanding.

The first idea is about *heuristics*. According to it, existing mechanistic explanations can serve as templates, schemes or recipes for the search of causes. The knowledge about a possible mechanism tells the researcher what to look for and where. This simplifies the search for causes, especially in situations where one can be confident that the menu of possible explanatory mechanisms covers all the plausible alternatives.

The second idea is related to *justification of causal claims*. It posits that the knowledge about possible mechanisms can provide support for causal claims. Causal claims without an account of the underlying mechanisms are possible and in principle fully legitimate, but knowledge of a mechanism makes them much more secure. This idea originates from everyday thinking, but it is also accepted in scientific contexts. Of course, the idea can sometimes be misleading: an imagined causal mechanism can give false credence to spurious causal claims.

The third idea is about the *presentation* of explanatory information. A mechanism scheme provides useful means for presenting and organizing explanatory information. A narrative form makes the explanatory information more digestible for humans, and mechanism schemes can be regarded as templates for such narratives. They outline the central features of the explanatory narrative and help people to focus on the right pieces of information.

The fourth idea concerns *the organization of social scientific knowledge*. According to an old empiricist view, general knowledge in science consists of empirical generalizations and more abstract theoretical principles from which these generalizations can (ideally) be deduced. The mechanistic account of knowledge challenges this picture on two counts. First, the locus of generality (and explanatory power) in social scientific knowledge is considered to lie in the mechanism schemes. The social sciences do not have that many valid empirical generalizations and those that they have are not very explanatory. On the contrary, they are among the things that require explanation (Cummins 2000). Explanatory power – and general applicability – comes from knowledge of possible causal mechanisms. When social scientific knowledge expands, it does not do so by formulating empirical generalizations that have broader application, but by adding or improving items in its toolbox of possible causal mechanisms. This brings us to

the second challenge to the traditional picture of social science knowledge: the ideal of knowledge in no longer an axiomatic system, but a much looser idea of an expanding theoretical toolbox. The expectation is that mature social science would be more like a textbook of modern cell biology than a treatise in elementary geometry.

These are important ideas – and I subscribe to all of them – but they do not address the issue of explanatory relevance. They do not tell us what is the explanatory import of mechanisms. Nor do they tell us how to construct a good or satisfactory mechanistic explanation. Furthermore, they provide no guidance for the evaluation of suggested mechanistic explanations.

The insufficiency of mechanistic ideas

To see why the idea of mechanisms is not very helpful in dealing with the problem of explanatory relevance, we have to understand how this notion works. Let us begin by distinguishing two ways to talk about mechanisms. Examples of both are plentiful in the literature and most of the time people seem to assume that they incorporate the same notion of mechanism.

The first (let us call it A-mechanism) regards mechanisms as *a particular causal chain*. The mechanism is whatever connects the cause and effect. No matter how long or complicated the causal process is, it can be called a mechanism if its description answers the question *how did the cause bring about the effect?* The second way (B-mechanism) to talk about a mechanism regards it as *a theoretical building block*. Here the mechanism is an object of theoretical interest, and it often applies only to simplified and idealized explananda. The above mechanistic idea about the organization of general knowledge is based on this notion of mechanism.

The differences between these two notions can be seen more easily if we consider their interrelations. The first thing to observe is that a single A-mechanism can involve many B-mechanisms. A number of different B-mechanisms can work serially or simultaneously in the same causal process. Furthermore, nothing prevents B-mechanisms from working in opposite directions. For example, if we are explaining the rise in the murder rate after the collapse of a corrupt central government, some of the B-mechanisms could work toward a reduction of crime, whereas others would increase it. The *explanandum* of an A-mechanism is the total (or net) effect, whereas the *explanandum* of a B-mechanism is more like a component effect. Clearly we should not use these two notions of mechanism interchangeably.

The idea of an A-mechanism builds upon the idea that the knowledge of the details of a causal process makes the explanation better. This idea feels intuitively correct. People prefer detailed explanations to mere sketches of explanation. This preference shows that detail is regarded as a virtue in explanatory contexts. The crucial question here is what do we mean by more detailed? An explanation is more detailed when it omits less of the relevant information. The key word here is relevant. The details of the causal process can be quite harmful for the explanation if they are irrelevant from the point of view of the explanandum. There has been some discussion in the philosophy of science about whether the addition of irrelevant details makes the explanation completely non-explanatory or whether it just makes an explanation worse (see Salmon 1998). We do not need to take a stance on this partly verbal issue; it is sufficient to point out that irrelevant details at least decrease the *cognitive salience* of the explanation. Irrelevant details can mislead: in such cases we identify wrong factors as explanatory. A more detailed causal story might also be more difficult to grasp. This is a simple fact of human cognition: our memory and ability to focus are limited and burdening them restricts our inferential performance. In both cases our ability to answer *what if* questions decreases. In other words, we will understand less.

Although the notion of an A-mechanism is important, it does not provide much insight into the nature of explanation. It is a placeholder notion: whatever explains the fact that *c* caused *e* is the mechanism. In this sense it is of limited analytical value: it names the challenge, but does not provide tools for dealing with the problem of explanatory relevance. Nor does the notion of an A-mechanism give any guidance for constructing mechanistic explanations. It does not tell us which level of organization we should focus on, which elements of the process should be incorporated into the mechanism, or how detailed the description of these items should be.

The identity criteria for B-mechanisms are stricter, but this notion faces the opposite problem. B-mechanisms are in many ways analogical to component causes. Just like component causes, they can be involved in cases of *overdetermination*, in which more than one mechanism works to guarantee a certain outcome. (Sometimes they could work simultaneously; sometimes one mechanism could pre-empt another.) Similarly, they could be involved in the cases of counteracting causes, in which the mechanisms compete by having opposite causal effects. Finally, as with an individual component cause, a B-mechanism can be simply explanatorily irrelevant. In the case of an A-mechanism the connection to the explanandum is guaranteed by definition, but such a guarantee

is lacking in the case of a B-mechanism. Any causal process involves many B-mechanisms at various levels of organization (physical, chemical, etc.); it is an open question whether they are in any way relevant to the explanandum we are interested in. Here again we find the problem of explanatory relevance: we need some guidance with B-mechanisms. What is the appropriate level of organization to focus on? What makes a given B-mechanism relevant to the explanandum? In how detailed a manner should we describe the mechanism?

When people construct and evaluate mechanistic explanations they are intuitively making judgments about explanatory relevance. Everybody agrees that good mechanistic explanations capture the relevant aspects of the causal process. However, nothing in the idea of a mechanism helps us in making and evaluating these judgments – we have to trust our intuitions. The trouble with intuitive assessments of explanations is that they are based on unreliable metacognitive processes. The sense of understanding is highly unreliable, and this shows in the disagreements people have in their judgments about explanatory value. Given the intuitive appeal of mechanistic storytelling in everyday reasoning, the critics of mechanistic theories of explanation might be right in being skeptical about their value. Especially when we are working with highly abstract sketches of mechanisms (as is typical in the social sciences), the danger of the illusion of depth of understanding is a real one.

Clearly, the notion of a mechanism is not a sufficient tool for improving explanatory practices in the social sciences. It needs to be supplemented with additional ideas from the theory of explanation. I will now turn to these ideas and try to show how they can be used to make sense of mechanistic explanation.

The importance of the explanandum

The problem of explanatory relevance in causal explanation is that the causal history of an event includes a vast number of elements and aspects that are not explanatorily relevant to the explanation-seeking question we are addressing. A natural way to start sorting out this problem is to take a closer look at the explanandum.

A contrastive account of explanation is helpful here. According to this view, explanations are answers to questions in the following form: why the fact rather than the foil happened? We want to know why things are one way rather than some other way. An anecdote about the famous bank robber Willie Sutton illustrates the basic idea of contrastive explanation. When Sutton was in prison, a journalist asked him why he

robbed banks. Sutton answered, "Well, that's where the money is." The journalist had in his mind the question: "Why do you rob banks, instead of leading an honest life?" whereas Sutton answered the question: "Why do you rob banks, rather than gas stations or grocery stores?" This is the basic insight of the contrastive approach. We do not explain simply "Why *f*?," rather, our explanations are answers to the contrastive question "Why *f* rather than *c*?" (Garfinkel 1981: 21–2). Instead of explaining plain facts, we are explaining contrastive facts (Ylikoski 2007).

Sometimes the contrast is imagined: we ask why an object has a particular property rather than a different property we expected it to have or our theory predicted it would have. Sometimes the contrast arises from a comparison: we ask why an object has a certain property rather than being like an otherwise similar object that has a different property. In both cases we are explaining a difference: why a fact is the case rather than its exclusive alternative.

I will adopt the convention of expressing the contrast in the following manner: *fact [foil]*, which should be read as fact rather than foil. The number of foils is not limited; there can be more than one. The contrastive idea does not put strict limitations on kinds of facts that can serve as the explananda: they can belong to different ontological categories. They can be properties, events, quantities or qualities. The crucial thing is that the fact and its foil should be exclusive alternatives to each other. To cover the wide variety of possible ontological categories, I will here simply speak about *variables*. This terminology does not commit us to any specific ontology, but allows us to make our points more generally and to avoid messy philosophical debates about the nature of events and other ontological categories. The basic idea is that whatever the real relata of explanation are, they can be represented as variables.

The idea of contrastive explanandum is a practical tool for analyzing explananda and for making the intended explanandum more explicit. Although it is not always apparent from the surface appearance of explanations, all explanation-seeking questions can be analyzed and further explicated by articulating the implicit contrastive structure of the explanandum. This explication makes the aims of explanation-seeking questions more clear and makes the evaluation of the adequacy of the proposed explanations easier. Articulating the contrastive structure forces one to be specific about the intended object of the explanation and about the respects in which we think the object could have been different (Ylikoski 2007). Are we explaining a singular event or a more general phenomenon or regularity? Are we addressing properties of individuals or of populations? What is the appropriate level of

description: are we after micro-level details or patterns found at the macro level?

Quite often the original question turns out to be a whole set of related contrastive questions. This is a good thing: *smaller questions are something that we can realistically hope to answer by means of empirical enquiry*. Contrastive articulation is also useful in controversies over apparently conflicting explanations. Often the apparently competing explanations turn out to be addressing complementary or completely independent questions. This is as it should be: we can be pluralists about explanation-seeking questions, but whether the answers are correct is still an objective matter.

The contrastive approach does not only help to clarify the explanandum; it can also be used to evaluate explanations. Rather than asking what the intended explanandum is, we can also ask *what is the contrast that the given explanation (or piece of explanatory information) can explain?* This is very useful approach, particularly, in social science controversies about explanation. Rather than getting entangled with difficult interpretive problems about what certain theorists are really attempting to explain, we can take a look at the explanations they provide and consider what they *in fact* explain (Ylikoski 2007).

The contrastive idea underlies our preferred form of causal inquiry: the scientific experiment. We contrast the control group with the experimental group or the process after the intervention with the process before the intervention. In both of these cases, we are trying to account for *the differences* between the outcomes. The basic idea is to keep the causal background constant and to bring about changes in the outcomes by carefully controlled interventions. A similar contrastive setup motivates comparative research. In general, the idea that explanations are contrastive is natural if one thinks that the aim of explanation is to trace relations of dependency. Our goal is to understand how changes in the cause variable bring about changes in the effect variables. We want to know what makes the difference and then leave out the factors that do not. The contrastive idea is based on a very intuitive feature of our explanatory cognition.

Making a difference

The starting point of the contrastive-counterfactual approach to explanation is realistic: explanations attempt to trace objective relations of dependence. These dependences can either be causal or constitutive. Dependencies are objective in the sense that they are independent of our ways of perceiving and theorizing about them. They are also

separate from our means of describing them. That is to say, the real relata of explanation are facts, not sentences.

If the explanandum is contrastive, as I have suggested, it is natural to think that the explanans is also contrastive. Following this idea, philosophers of causation are increasingly thinking that causation is doubly contrastive (Schaffer 2005; Woodward 2003). The same idea applies also to causal explanation. In the simplest form, causal explanatory claims have the following structure:

c [c]* explains *e [e*]*

In plain language, the cause variable having the value *c* rather than $c^\star$ explains why the *explanandum* variable has the value *e* rather than $e^\star$. A natural way to understand this relation is to regard it as *a counterfactual dependence*: if the cause variable had had the value $c^\star$, the value of the *explanandum* variable would have been $e^\star$ rather than *e*. Note that although the claim is about counterfactual situation, *there is nothing counter to the facts in the relation of dependence itself*. Also of importance is that counterfactual dependence is a modal notion: *explanation is not about subsumption under empirical regularities, but about counterfactual dependence*.

The idea of a counterfactual theory of causation is very old. We can avoid some of the historical problems of counterfactual theories of causation if we do not regard it as a reductive theory of causation and if we limit its application to causal explanation. The new idea is to combine it with the idea of contrastive explanandum. Together with a sophisticated manipulation account of causation this idea helps us to solve most of the problems that counterfactual theories of causation have faced (see Ylikoski 2001).

An important problem for the counterfactual theories has been the specification of the truth conditions for the counterfactuals. A manipulation account of causation is helpful here. In the manipulation account of causation

c [c]* causes *e [e*]* if we can bring about $e^\star$ *[e]* by bringing about $c^\star$ *[c]*

This cannot serve as a reductive analysis of causation, as "bringing about" is already a causal notion. However, it can serve as a starting point for an adequate descriptive analysis of the notion of causation. An advantage of the manipulation account is that it provides a natural way of distinguishing between *real causal relations* and *mere correlations*. Real causal relations can be used as bases for effective interventions, whereas mere correlations do not allow this.

Experiments and other causal interventions in the real world are often quite messy. To make sense of the meaning of causal claims we need the

1. *I* does not change *Y* directly
2. *I* does not change the value of any causal intermediate *S* between *X* and *Y* except by changing the value of *X*
3. *I* is not correlated with some other variable *C* that is a cause of *Y*
4. *I* acts as a switch that controls the value of *X* irrespective of *X*'s other causes *U*.

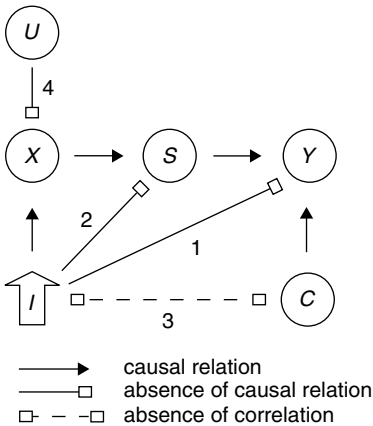


Figure 8.1 Causal relations and ideal intervention.

notion of an ideal intervention, developed by Woodward (2003: 94–151). An ideal intervention *I* changes the value of effect variable *Y only via* changing the value of cause variable *X* as shown in Figure 8.1 (The formulation and the graph are adapted from Craver 2007.)

The notion of ideal intervention gives us straightforward semantics for causal counterfactuals. It also makes it possible to avoid anthropocentric features of earlier manipulation accounts of causation: the definition does not refer to a human agent. Nor must we assume that the intervention is humanly possible – we are only interested in possible interventions. For us the main advantage of this account is that it provides us with a natural way of making sense of counterfactuals in causal explanation, meaning that an explanatory claim is correct if ideal intervention on the explanans variable had brought about the appropriate change in the explanandum variable.

An interesting feature of Woodward's account is his view of explanatory generalizations about effects of ideal interventions (Woodward 2003: 239–314). These *invariances* lack many traditional features of laws of nature. They usually hold only for a limited range of possible interventions (and changes in background factors); they can refer to particular objects, places and times; and they might contain exceptions. In short, they might hold only in a certain domain and break up outside of it. However, in contrast to the traditional empiricist accounts of laws of nature, invariance is a modal notion: it makes a claim that the invariant relationship between cause and effect variables will hold under a set of possible interventions. According to Woodward, invariances capture

what is essential from the point of view of explanation, whereas most of the traditional attributes are superficial from this angle. Consequently, in his account all explanatory generalizations are not laws. This is good news for the social sciences (and other sciences outside of fundamental physics): the generalizations satisfying the traditional criteria of lawhood are rare or nonexistent.

It is plausible to think that our contrastive explanatory preferences stem from our nature as active interveners in natural, psychological and social processes. We desire to know where to intervene to produce the changes we want, and this knowledge often presupposes answers to some *why* and *how* questions. Without this knowledge one would not know when the circumstances are suitable for an intervention and one would not be able to predict the results of the intervention. I do not wish to claim that the notion of explanation can be reduced to its origins in agency. After all, we are also interested in explaining things that cannot be humanly manipulated. However, our instrumental orientation might still explain why our explanatory preferences are as they are.

The explanatory import of mechanisms

In the contrastive counterfactual account causal claims are based on change-relating invariances, meaning that no mechanism is needed for making a causal claim. This helps us to avoid the regress problem that an account which assumes that mechanisms are an integral part of all causal claims would face. However, it does not challenge the legitimacy of the assumption that outside of fundamental physics there will always be a causal mechanism mediating between the variables. Nor does it imply that we should not try to have knowledge about the mechanisms (Hedström and Ylikoski 2010).

However, there is one important consequence: macro variables are explanatory in the same sense as micro variables. In other words, the explanatory factors (for explananda at individual or social levels) might be found at the social level. The crucial issue is whether the right kind of invariance exits between the explanans and the explanandum variables. If by changing the macro-variable we can bring about a relevant change in the explanandum variable, we have identified the cause. The intervention account does not privilege any specific level of description; it is a purely empirical matter which level provides the variables that inform us about the relevant invariances (Steel 2006). The crucial factor is the contrast on the explanandum side: depending on the contrast, a micro or macro variable might turn out to be more relevant (Ylikoski 2001: 77–100). The relevance is determined by the nature

of invariances that the variables are involved in. This blocks the simple argument for methodological individualism through causal mechanisms. Causal claims do not have to incorporate mechanisms and they can refer to macro variables. Clearly, if we want to stick with methodological individualism, we need some independent arguments.

If causal claims can be made without referring to mechanisms, the question arises, what is their function in causal explanation? Is it a misguided idea that adds nothing to the process of causal explanation, as some critics have suggested? This would be an overreaction. It is important to recognize the limited scope of the above claims. These are claims about singular causal statements that answer to simple explanation-seeking questions. However, when we turn our attention to the broader issue of understanding, the true contribution of mechanisms becomes apparent. A singular causal claim might be based on a simple claim about counterfactual dependence, but for theoretical understanding we need more.

In the erotetic approach, explanation-seeking questions are not treated separately as independent entities. The questions come in clusters of closely related *what if* questions. The individual questions are related to each other in many ways. When we find that the value of variable Y is dependent on the value of variable X in a manner that allows (at least an ideal) intervention, this raises a whole series of questions. First, *why does the counterfactual dependence hold?* Second, *what is the range of interventions that this invariance allows without breaking apart?* Third, *what are the background conditions for this invariance* and *how sensitive is the invariance to changes in these conditions?* Fourth, *are there other alternative interventions that can bring about the same changes in Y?* Fifth, *is it possible to account for a more precise explanandum?* And sixth, *is it possible to find a generalization that stays invariant in a broader range of interventions and background conditions?*

The knowledge of mechanisms is involved in answering all these questions. The questions are related to each other via *chains of presupposition*. As we can see from the first question, underlying every *why* question is a possible *how* question. Individual claims about causal influence are of limited interest. We want to know why X can cause changes in Y. Answering this question requires knowledge about causal mechanisms: we have to know how the changes in X are transmitted to changes in Y. The fourth question about alternative interventions is closely related and knowledge about possible mechanisms greatly facilitates answering it. Answers to both of these questions expand the range of *what if* questions we can answer about the phenomenon – they increase our understanding.

An important manner in which knowledge about mechanisms advances our understanding is *integration*. An answer to the first question integrates the causal claim with other pieces of knowledge. This is a very important consideration from the point of view of understanding. When an explanation is integrated into a larger theoretical framework, the theoretical connections can expand the range of answers to different *what if* questions in two ways. First, the integration allows for inferential connections to an already existing body of knowledge, and this might make it possible to find unforeseen dimensions in which *what if* questions concerning the explanandum phenomenon can be answered. Second, the explanation itself may bridge previous gaps within the existing theory and thus enable answers to new *what if* questions that do not directly concern the original explanandum phenomenon (Ylikoski and Kuorikoski 2010).

The second and third questions arise from the fact that we want to know how broadly our causal knowledge can be exploited; in other words, we wish to know how many other *what if* questions we can answer with the same information. When we know more about the mechanism transmitting the causal influence, we have a better idea of what kinds of factor can disrupt the causal link and how. With this information, we can answer more *what if* questions concerning situations in which the background assumptions or conditions are different, for example in situations where some parts of the mechanism were altered or when the background factors not included in the explanatory generalization change. In this way, knowledge of mechanisms contributes to understanding the domain of application of the knowledge about the invariance. Without knowledge about the mechanisms we would have trouble evaluating the range of *what if* questions that can be answered with our piece of causal knowledge. Knowing the limits of one's knowledge is an important ingredient of understanding.

Finally, the fifth and sixth questions suggest that knowledge about the mechanisms might help us to improve the explanatory generalization. A common way to search for more general formulations is to look at the underlying mechanisms and explore the possibility of integrating them into the explanatory generalization. This might make it possible to formulate the explanatory invariance in such a manner that it allows us to explain either *a sharper explananda* or *a broader range of explananda* than the original (Ylikoski and Kuorikoski 2010). It might also make it possible to formulate a more general statement of invariance that holds for a broader set of interventions and/or background conditions. More general invariances are preferable, as they also make it possible to answer a broader set of *what if* questions. In both of these cases the road

to better explanatory generalizations goes through mechanisms: they suggest variables that could be incorporated into the explanatory generalization (Ylikoski and Kuorikoski 2010).

If these observations are correct, the call for mechanisms really makes sense. Knowledge about mechanisms means an ability to answer more *what if* questions, i.e. more understanding. When this explanatory contribution is combined with the four other functions of mechanisms mentioned earlier, we see that the social mechanisms movement has been motivated by the right kind of intuitions. However, it might be that some of the advocates of social mechanisms have had the wrong ideas about the source of legitimacy of these intuitions: the necessity of mechanisms does not derive from ontological arguments for methodological individualism or from semantics of individual causal claims.

Conclusion

In this chapter I have suggested that the current mechanistic ideas about explanation are insufficient for the purposes of analytical social science. The aim is to improve explanatory practices, but the notion of a mechanism is not very helpful. I have also suggested a number of ideas about explanation that can be used in improving social science explanations. The first is the erotetic approach to explanations combined with the idea that all explanation is contrastive. This idea is useful in making explanations more explicit and by allowing one to be more precise about the intended explanandum, as it permits a much sharper evaluation of explanatory claims. It also makes it possible to think about interrelations between explanations by considering chains of presupposition. The second idea is the counterfactual criterion of explanatory relevance. When explanatory claims are understood as claims about possible (ideal) interventions, we get a notion of explanatory relevance that is both intuitive and powerful. Among other things, this approach allows for a novel way of understanding explanatory generalizations as claims about invariances. Third, I have suggested that we employ the notion of understanding to make sense of the contribution of the knowledge about mechanisms to our explanatory enterprise. A singular causal claim does not require knowledge about the mechanisms, but once we begin to consider the broader set of *what if* questions, the importance of mechanisms becomes apparent. The notion of understanding also helps us to see the unreliability of the intuitive way of evaluating the relevance of explanatory information.

As the sense of understanding can be highly misleading, we have a further motivation for seeking a more explicit theory of explanation. Finally, I have suggested that we should decouple the idea of mechanistic explanation from the doctrine of methodological individualism. These two ideas should be argued separately, not as grounds for each other.

REFERENCES

Achinstein, Peter. 1983. *The Nature of Explanation*. Oxford University Press.

Craver, Carl. 2007. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Clarendon Press.

Cummins, Robert. 2000. "'How does it work?' versus 'what are the laws?': two conceptions of psychological explanation," in F.C. Keil and R.A. Wilson (eds.), *Explanation and Cognition*. Cambridge: MIT Press, 117–44.

Garfinkel, Alan. 1981. *Forms of Explanation*. New Haven: Yale University Press.

Gopnik, Alison. 2000. "Explanation as orgasm and the drive for causal knowledge: the function, evolution, and phenomenology of the theory formation system," in Keil and Wilson (eds.), *Explanation and Cognition*. Cambridge: MIT Press, 299–324.

Hedström, Peter and Petri Ylikoski. 2010. "Causal mechanisms in the social sciences," *Annual Review of Sociology* 36: 49–67.

Hempel, Carl. 1965. *Aspects of Scientific Explanation*. New York: The Free Press.

Hesslow, Germund. 1983. "Explaining differences and weighting causes," *Theoria* 49: 87–111.

Hitchcock, Christopher C. 1995. "Discussion: Salmon on explanatory relevance," *Philosophy of Science* 62: 304–20.

Rozenblit, Leonid and Frank C. Keil. 2002. "The misunderstood limits of folk science: an illusion of explanatory depth," *Cognitive Science* 26: 521–62.

Salmon, Wesley. 1998. *Causality and Explanation*. Oxford University Press.

Schaffer, Jonathan. 2005. "Contrastive causation," *The Philosophical Review* 114: 297–328.

Schwitzgebel, Eric. 1999. "Children's theories and the drive to explain," *Science & Education* 8: 457–88.

Steel, Daniel. 2006. "Methodological individualism, explanation, and invariance," *Philosophy of the Social Sciences* 36: 440–63.

van Fraassen, Bas. 1980. *Scientific Image*. Oxford University Press.

Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. Oxford: Basil Blackwell.

Woodward, James. 2003. *Making Things Happen. A Theory of Causal Explanation*. Oxford University Press.

Ylikoski, Petri. 2001. *Understanding Interests and Causal Explanation*. PhD thesis. University of Helsinki, May 2001. Available at: http://ethesis.helsinki.fi/julkaisut/val/kayta/vk/ylikoski.

2007. "The idea of contrastive *Explanandum*," in J. Persson and P. Ylikoski (eds.), *Rethinking Explanation*. Dordrecht: Springer, 27–42.
2009. "The Illusion of Depth of Understanding in Science," in H. de Regt, S. Leonelli and K. Eigner (eds.) *Scientific Understanding: Philosophical Perspectives*. Pittsburgh University Press, 100–119.
Ylikoski, Petri and Jaakko Kuorikoski. 2010. "Dissecting explanatory power," *Philosophical Studies* 148: 201–19.

9 Causal regularities, action and explanation

Pierre Demeulenaere

> A prisoner who has neither money nor interest, discovers the impossibility of his escape, as well when he considers the obstinacy of the gaoler, as the walls and bars with which he is surrounded; and, in all attempts for his freedom, chooses rather to work upon the stone and iron of the one, than upon the inflexible nature of the other. The same prisoner, when conducted to the scaffold, foresees his death as certainly from the constancy and fidelity of his guards, as from the operation of the axe or wheel. His mind runs along a certain train of ideas: The refusal of the soldiers to consent to his escape; the action of the executioner; the separation of the head and the body; bleeding, convulsive motions, and death; but the mind feels no differences between them in passing from one link to another.
>
> (David Hume [1758] 1975: 90)

Nineteenth-century social studies are divided over a major issue: are they similar to the study of nature, introducing laws governing individual behaviors and social outcomes; or are they intrinsically different, since there are no major regularities and laws to be found in social life? Mill (1843) argued that there was an underlying similarity between the social and the natural sciences. Dilthey (1883) by contrast declared them to be fundamentally heterogeneous, the social world being mainly historical, to be apprehended through the description of particular events, the sciences of nature being characterized by their orientation to general laws. It was Windelband ([1894] 2000) who coined the terms by which the two approaches would be characterized and contrasted: nomothetic sciences being directed to the determination of general laws, and idiographic sciences describing singular events (Abbott 2001).

Yet Windelband actually never said that there were two fundamentally different kinds of science, forever ontologically in opposition; he only claimed that any object, natural or social, could be approached from either a nomothetic perspective, or from an idiographic perspective. Any object (a given language for instance) displays regular as well as historical features; we should not therefore contrast two kinds of science, but assume that any particular science confronted with given

objects will proceed to deal with these objects both nomothetically and idiographically.

Hence the social sciences should be considered neither uniquely nomothetic nor uniquely idiographic. Such a clear-cut exclusive distinction is not appropriate to any science. Every science involves both the description of particular and historical events that cannot be deduced from general laws, and the characterization of general regularities that govern large groups of particular facts. Of course, not all sciences involve the same range of nomothetic evidence: history is not physics, and it is clear that physics involves much more nomic regularity than history. Nevertheless, the physical world does include particular events that are not subsumed by general laws explaining the course taken by such events; for instance, mutations in the genetic world. There are "novelties" in the physical world that cannot be subsumed under known general laws. Similarly, there are regularities in the historical world that allow us to "explain" some phenomena by referring them to more general patterns of action. It is true that those patterns do not always correspond to the exact idea of a natural law; and we will return below to this important point. However, "explaining" an action or a behavior necessarily invokes the kind of regularity which allows us to speak of explanation. Without such regularity the notion of "explanation" is meaningless. It remains nonetheless necessary to clarify how explanation works in the social sciences, and on what type of regularity it might rest.

Two opposed conceptions should therefore be treated as irrelevant for the social sciences. First, we should rule out the complete reduction of historical variety to the deduction of events on the basis of general laws. It should be acknowledged that many events cannot be reduced to general laws, and are simply described as singular transient phenomena, description of which is already an important and informative task for the social sciences: for instance, what exactly are the kinship rules that govern a given society? This is a matter of description and has its own difficulties and challenges; it is not necessarily possible to reduce the particularities that are thereby disclosed to general laws on which they are dependent. Carl Menger ([1883] 1996) excluded all historical considerations in the development of "pure," timeless economic laws, from which historical events could then be deduced. Such a view seems today outdated; time does matter, and we cannot straightforwardly refer to economic "laws" without taking into consideration context and cultural situations.

On the other hand, it does seem necessary to reject any claim that the social sciences are inherently particular and descriptive, the events which

are studied being irreducible to any general approach based on regularities and laws. Such a view, long after Dilthey and the *Methodenstreit*, is defended by authors such as Passeron (2006) or Veyne (2008), for example. The latter derives his position from that of Foucault: history and the social sciences can only describe particular events; they are irreducible to and do not depend upon any kind of generality. This extreme position would ultimately imply that no "explanation" is available to history or sociology at all, since the answer to the question "why?" implicitly involves some kind of regularity that allows us to say "because." Max Weber is for instance well-known for his criticism of the idea that the social sciences should be directed to the production of general laws. He insisted that social science is always concerned with particular events and their historical significance. This is the position for which he is usually remembered. But, conversely, he never claimed that explanation in the social sciences placed no reliance at all upon general laws. On the contrary, he always maintained that, while the aim of the social sciences is not to reduce particular events to general laws, general laws are nevertheless necessary to illuminate given historical and particular connections. This position, taken from Rickert (1902), is more nuanced, and cannot be reduced, as in Veyne and Passeron, to an emphasis upon strictly particular events only interpretable in terms of ideal-types, despite their claim to be Weber's followers. Weber repeatedly reaffirmed the importance of laws and causal connections for the explanation of particular events (Weber [1904] 1991).

It should be added that an exclusive emphasis on particular description would ultimately lead to the rejection of the very idea of description, since any description involves the possibility of translation into a given language of situations often alien to it. It implies some kind of regularity that permits translation (a point denied by some philosophers, notably Quine). But more specifically, quite often description does in fact presuppose the identification of *causal regularities*, beyond mere regularities, if it is to be intelligible. I will take a very simple example from Geertz to illustrate this:

> The French [the informant said] had only just arrived. They set up twenty or so small forts between here, the town, and the Marmusha area up in the middle of the mountains, placing them on promontories so they could survey the countryside. But for all this they couldn't guarantee safety, especially at night, so although the *mezrag*, trade-pact system was supposed to be legally abolished it in fact continued as before. (Geertz 1973: 7)

There are two basic explanations in this passage that go beyond sheer description. First, why are the forts placed on promontories? So that the French could survey the countryside; this explanation is based on

a general and practical causal regularity. Second, why could they not guarantee safety, especially at night? Because mutual retaliation continued to prevail, although supposedly abolished; moreover, it is difficult to monitor people in the mountains, especially at night. Together, the incidence of an institution on behaviors is implied, combined with the practical difficulty of observing people at night. So it can be said that such descriptions do in fact involve explanations based on general causal action regularities. We are not in a situation of complete indeterminacy of description, because the regularity of action is obvious.

In this context, the notion of mechanism is a challenging and difficult one:

1. On the one hand, it clearly implies some kind of regularity, causality and predictability; in the absence of such features it would seem an abuse of language to speak of a mechanism. Clearly, the notion stems from situations in the physical world where such law-governed regularities do exist. An engine, for instance, is a mechanism, since it displays law-governed regularities that trigger its functioning. "Mechanisms" are often introduced rhetorically in social science literature without any strong evidence that such things really exist, no evidence of strong causal regular connections being given.
2. On the other hand, it has to be linked to the historical aspect of social life, and to the necessity of taking into account the uniqueness of events, the very frequent difficulty of making predictions, and the apparent difficulty of using the notion of "law" in the social sciences. The intervention of a "mechanism" is often linked to the alleged difficulty or impossibility of finding laws in the social world. It appears to be an alternative explanatory strategy in a social world for which no laws are available (Elster 1989).

Surprisingly, the recent literature on mechanisms, which is certainly not homogeneous, tends to distance itself from the idea of law (for an overview see Hedström 2005), but not from the ideal of causality. The literature is therefore ambiguous about the very notion of mechanism; it does seem meaningless to introduce the existence of a mechanism without any implication of causal determination; and a causal determination seems necessarily to depend upon something like a law, or at least a causal regularity. Referral to a mechanism would otherwise be only a verbal proof, lacking underlying evidence. But since the notion of mechanism is introduced precisely because of the difficulty of discerning laws in the social world, the consistency of the argument may well seem dubious.

Moreover, Dilthey's attack upon the reduction of social sciences to the program of nomothethic natural sciences presumed that the cultural

sciences depended upon the "understanding" (*Verstehen*) of the "meaning" of intentional actions taking place in cultural settings, as opposed to the natural sciences where no such meaning and understanding is involved. It is therefore interesting to note a revival of this emphasis upon intentions and meaning through mechanism-based explanation, joined to a critique of covering-law explanations which are criticized for their neglect of the meaning of intentions.

I therefore wish to address in this chapter the question of understanding the kind of causal regularity that is to be found in social sciences explanations. I will proceed as follows. In the first section I will describe typical features of ordinary actions, stressing the fact that they depend, for their successful execution, upon predictable causal regularities. I will then in the second section seek to show that social science explanations rest on the same kind of causal regularities as those necessary for ordinary action. I will conclude by criticizing the idea that mechanism-based explanations should be strongly contrasted to covering-law explanations, since they both similarly depend upon the same kind of regularities. I will nevertheless argue that regularities in the social world should not always be considered to be laws, given the importance, in causal chains, of rules, norms and institutions that cannot be assimilated to laws, although they do play an effective role in causal sequences and hence in causal explanations.

The analytical dimension of ordinary action and causality

I will begin by emphasizing a few prime characteristics of ordinary action involving causality. I will thereafter assume that the features of ordinary action are the basis for social scientific causal explanations. There is however an important difference, in that the social sciences mostly intervene *ex post* with respect to facts, whereas action, at the point of decision, faces an unknown future which it apprehends *ex ante*. All the same, it is based upon former experience and previously discovered regularities. It does not exclude the possibility of sound predictions in cerain circumstances.

The first point to which I would like to draw attention is that a major feature of action in a social world is that it has regular features. There is no such thing as an action without strong regularities. Of course, there are not only regularities to be found; there is also a considerable degree of uncertainty and unpredictability in the realm of action, features often stressed for instance by Austrian economists such as von Mises, Hayek or Shackle (1972). Nevertheless, we cannot do without the idea

that any action presupposes foreseen causal regularities. This is in particular the core idea of instrumental action: the choice of certain means to reach an end rests on the regular fact that such means do regularly lead to such ends. The same can be said of an action that is oriented by values. It is based on the stability (at a certain time) of the link between values and the specific actions which implement them. For example, faithful Catholics are not supposed to abort pregnancies. We expect therefore that they will tend not to have abortions, so that they might be consistent with their Catholic faith. Here we have something fundamentally similar to the case of instrumental action. Many Catholics do not have abortions so that they can be consistent with their faith, and so implement the values in which they believe; they would otherwise be confronted with cognitive dissonance (which naturally often occurs). The action is based on regularity: Catholicism proscribes abortion, and the action of not aborting pregnancies reinforces conformity with the faith. This is the central aspect of intentional action, which is always oriented to the predictable outcome of an action. Action depends on the fact that predictable results are brought about when certain activities corresponding to certain intentions are set in motion.

We can ideally distinguish three possible cases where this occurs:

1. When an individual acts on inanimate nature the outcome is usually predictable, determined by the operation of natural laws: for instance, agricultural technology (crop rotation and fallow for example) rests on natural regularities which are progressively discovered and can thereafter be foreseen. These regularities can be apprehended at different levels, whether practical or chemical in this case. It was gradually established that the regular planting of legumes (like peas or beans) increased the fertility of soils. Long afterwards it was discovered that this was because legumes and grains are complementary: the one returns to the soil the nutrients that the other has used and removed from the soil. Practical discovery preceded scientific explanation. It allowed people to abandon the practice of leaving land fallow. Similarly, human interaction with animals rests largely on predictable regularities: the way animals are bred (horses and dogs for instance) presupposes natural regularities that allow human action to control many aspects of animal life.
2. When an action is isolated, not directly involving nature, animals or others, it depends on the independent features of such actions: for example, swimming depends on the performance of certain typical movements in the absence of which we would drown. (However, in an example like this action clearly depends on natural laws.)

3. Things become more complicated when interaction between human individuals is involved. This is the principal case for the understanding of social situations. In economics, the hitherto prevailing notion of instrumental action has been replaced by the ordering of preferences, which is less obviously instrumental. What was previously thought of in terms of instrumental action (in particular involving exchange) did involve other parties: it was not an isolated action.

There are two major (ideally distinct) cases when such an interaction occurs.

3a. The first occurs when human interaction is governed by stable institutions, or rules that permit, under normal circumstances, the prediction of some regular outcome. For instance, if I murder a neighbour, and if my crime is discovered, and if I cannot plead legitimate self-defense or involvement in euthanasia, in a society in which criminals are not exposed to private retaliation I would normally be taken to jail. In the United States, I might be sentenced to death. This is a very predictable outcome. There are innumerable examples of such situations, which rest on two elements: some rules are commonly given and admitted in a society; and most people, in common circumstances, are expected to act according to their roles in respect of these rules; they apply them on the basis of globally stable and predictable intentions. The regularity rests both on the fact that rules exist, and that most people tend to follow them. In such circumstances it is common to speak of "institutional" mechanisms. They clearly involve causal links: my crime will, under specific but typical circumstances, "cause" my condemnation.

3b. Even when we are not in situations where behavior is governed by stable rules and institutions commanding general assent, interaction often produces regular outcomes because behaviors are predictable. It is here more difficult to find simple examples, since most human interaction is linked to social norms and institutions; we never act in a pure state of nature. Nevertheless, some situations not directly governed by norms and institutions do tend to have predictable outcomes: for instance, the more I go to art galleries, the more I will know about painting, the more expert I will be in determining what kind of painting I am looking at. In addition to that, the more complex my judgment will be compared to that of someone who knows nothing about painting (in the same way that I know nothing about sumo wrestling for instance), and the more difficult it will be for me to share the judgments of those who

> are not really interested in painting, but who nevertheless express value judgments in respect of painting. Of course, this example relies heavily upon institutions and norms; but the main point is that the refinement of judgment based on wider knowledge does not in itself derive from norms and institutions. Consequently, it is more or less predictable that differences of knowledge in any given artistic field will typically "produce" a range of different judgments, reactions and preferences. We are therefore frequently engaged in actions where we anticipate others' reactions, not on the basis of specific institutions or rules, but on the basis of typical expected behavior under typical circumstances. For instance, if I lie many times to someone who eventually finds me out, it is only to be expected that this person will no longer trust me. Such an expectation is based on behavioral causal regularity. It leads us to anticipate the intentions that, in typical circumstances, will be caused in others. In ordinary life, our course of action is partially oriented by our knowledge and prediction of causally determined intentions that, under typical circumstances, arise in others. Clearly we could speak here of "psychological," or in a certain sense of "cognitive" factors, which would include rational behaviors and emotional behaviors.

In addition, it is important to note that even when social intercourse is organized through rules and institutions, their viability necessarily depends on typical attitudes which are arguably exogenous to those rules; I will return below to this important and difficult point. Two alternative views are defended in the literature of social sciences. The first treats the discovery of such rules and institutions that govern social behavior to be the major and ultimate task of social sciences, no further explanatory enterprise being attainable; this is a Wittgensteinian position (Descombes 1996; Winch 1958). I myself support the second, seeking understanding of the typically transcultural motives upon which such rules and institutions are based as and when they emerge, vary and are followed. Such a form of explanation involves causal mechanisms.

The difficulty thereafter is to distinguish what in such an explanatory effort should be considered to be an explanandum, and what the explanans. I have often contended (Demeulenaere 2008) that "costs" and "benefits" cannot be the ultimate basis on which to build all explanation of human behavior, since the estimation of costs and benefits often varies from one individual to another (or from one group to another: for instance regarding the negative impact of smoking). It is therefore more an explanandum to be explored and elucidated than

a straightforward explanans. Widespread attempts by economists to explain the emergence of institutions such as property rights in terms of costs and benefits (for instance Demsetz 1967; Eggertsson 1990) therefore encounter skepticism and critique on the part of anthropologists, since they consider the explanans (an individual estimation of personal costs and benefits defined in specific and restricted terms) to be the explanandum. We can certainly construct models in Prisoner's Dilemma situations where norms tend to emerge (Coleman 1990) on the basis of given utility functions; but utility functions in such situations do themselves depend on various social preferences (Hollis 1994; Sen 1977).

As a preliminary conclusion, I would like to emphasize the very great importance of such causal regularities in human social action, either based on rules and institutions, or dependant on typical transcultural motives and attitudes, together with practical constraints. Without such regularities social life would be extremely difficult, and barely manageable. Social life is, contrary to widespread opinion, overwhelmingly predictable; and our life is frequently driven by routine (Mantzavinos 2001). Such regularities can also be observed in our intercourse with nature, and in our relationships with others; these are based on the regular behavior of natural phenomena as well as on the regular behavior of individuals. The latter intervene either because they are stable institutions that connect peoples' decisions, or because there are typical reactions and motives independently of any type of institution which are regularly triggered and expected in certain circumstances. Those are moreover the elements upon which rules and institutions are built up (and thereafter explained). Action is dependent on "cognitive" (or "psychological") constraints that are quite stable, and on practical constraints (the way things can be done) discovered through experience in different circumstances. Most often the three aspects (psychological constraints, practical constraints and institutional constraints) are so deeply connected that it is difficult to separate the different dimensions. Institutional regularity is achieved because there are stable and predictable intentions underlying these institutions; for instance, people get married partly in order to provide order for the life of their children, and this is a fairly common feature of human life. Conversely, stable and predictable intentions exist partly because stable institutions exist, and it is understandable that people tend to get married because something like the institution of marriage exists, with its own features in a specific society, features which are not all necessary for the protection of children.

Nevertheless, it is obvious that causal regularity is not the only aspect of human action; irregularity and unpredictability are similarly

important. But ordinary individual and social actions necessarily rest on causal regularity and predictability. Four elements should be mentioned here regarding this relationship between regularity and irregularity in individual and social action.

First, any action necessarily rests on a *descriptive* basis, which means that we progressively discover, by encountering cognitive, practical and institutional constraints, the causal regularities on which our action is based. Such regularities depend on the way in which nature is organized and can be manipulated, how artifacts work, how other people typically react in certain circumstances, and how rules and institutions are established in a definite social situation. Our action is based on our descriptive knowledge of the architecture of natural and social situations and institutions, a knowledge which includes causal connections we discover in course of our intercourse with nature, artifacts and others. In respect of our relationships with others, we progressively discover the way they behave in typical circumstances, and the way that rules tend to coordinate individual intentions. In this context, instrumental action is a general concept which refers to various local features of actions. The relevant information necessary to engage in a definite action derives from the details of such configurations; the notion of instrumental action by itself does not tell us what are the means to achieve an end, we need instead specific empirical information about which specific means causally lead to specific ends, and we often depend on an institutional knowledge and framework for such information. To achieve our ends we need to know the institutions, and how other people defend them in particular ways. In the case of actions based on values, or on institutions, information is needed regarding what those values exactly imply, and which rules institutions include in their functioning. We have to be aware of value connections as well as institutional codes in order to be efficient in our undertakings. It should also be said that, very often, people act on false assumptions: they believe that particular actions will produce particular results, without any real evidence for such a connection. Their action is based on inaccurate, or false, descriptions.

Second, action proceeds necessarily on an *analytical* basis. Since it does not take account of all the elements that could obstruct the expected outcome, and since any course of action does in effect depend on many other elements, known and unknown, predictable and unpredictable, a decision to act analytically distinguishes relevant and fundamental aspects that are required for the achievement of the aim. It separates one causal link from all others. In the course of an ordinary action it is not usual to take into account the entire

range of contextual factors in the absence of which no action is at all possible (for instance, the overall role of the state, the laws of gravity, and so on). Any one action depends on a plethora of other dimensions to which it is connected and which permits its realization; but not all of them have to be considered in respect of the intended action. The analytical dimension is therefore essential to any decision to act. It isolates specific and more or less predictable sequences of cause and effect (whether or not mediated by institutions), disregarding all the other chains in which those local sequences are inevitably embedded. And so action focuses on the immediate elements that allow an aim to be achieved, without direct regard to the innumerable other conditions necessary for its fulfilment and which are assumed to be relatively stable. If we do not know the predictable causal consequences of our actions we are placed in a very difficult position for decision-making and action.

Third, human action in a social context is therefore always engaged on the basis of a *ceteris paribus* clause, implying that it is *abstract*. Actors first envisage effective actions on an abstract basis. An individual actor does not, and cannot, ordinarily take into account all the possible events which could endanger and interfere with the normal realization of his plans, resulting in an outcome different from the predictable result aimed at when he forms his intention of acting. In the previous paragraph, I emphasized that action focuses on the directly relevant means of realization of an aim, without considering the details of all the other necessary conditions which support the effectiveness of the means. Here I note the fact that, in addition to this analytical approach, there is also an abstract approach which, given a decision to act, discounts all those possible factors that might intervene and prevent the realization of the action. We certainly know that if we do not pay our rent, our landlord might be kind enough to let it pass, and leave us in our apartment in peace. This is a possible option, but not a plausible one. It is therefore ruled out in the course of ordinary action, for ordinary action is based on the belief that, *ceteris paribus*, if someone does not pay his rent, he will be sued by his landlord. In this case, there are two possible sources of irregularity that cut across the probable regularity that one expects to occur. Either there are other unanticipated events which will prevent the predictable event occurring: my landlord might die, and his heirs might become embroiled in a quarrel that prevents them from suing me in the short run; or, more unlikely, the landlord will not behave in the way he is expected to, for his own particular reason (he might suddenly decide to be a philanthropist). There are two distinct sources of irregularity:

1. either people do not behave in the way they were expected to, for whatever reason, unrelated to a change in external circumstances;
2. or new factors in the environment provoke a disturbance in ordinary behavior.

So when causal regularity is the most likely expected situation it could still be compromised by unforeseen factors. However, ordinary individual and social action does not take systematically into account all those possible dimensions which would prohibit anyone from acting. When an effective action is undertaken it is based on foreseen consequences founded on actual knowledge of causal links.

Fourth, the regularity of the outcome of an action is more or less predictable; it is more or less certain. For instance, the probability of exhausting a given piece of land would have been close to one in the Middle Ages if intensive cultures without fallow had developed; just as today red tuna overfishing will lead, in the absence of preventive action, to the exhaustion of the resource. But we are very often somewhat unsure of the outcome. For instance, we buy an apartment in a period when prices are increasing, interest rates are low, so that the demand for apartments is much greater than the supply. We can therefore anticipate that prices will continue to increase; but, as everyone knows, this is not necessarily the case, since other new elements could intervene that will make the prices fall. Here we are in a situation similar to that of physical events, meteorological events for instance. Some events are more likely than others, but in most circumstances we do not know what the exact probability is. Elster (2007) has emphasized the fact that sometimes two opposite outcomes can be triggered from a given situation. Simmel (1923) had earlier noted that a virtuous action can give rise to either gratitude or resentment. We are in fact in a situation similar to many others: we know that a good action will provoke either gratefulness or resentment, but we do not know what the exact probability of one or the other outcomes is. We know that normally it will not produce indifference. If more information is given, greater certainty can be achieved: if we know that a person is ordinarily kind and grateful, a good action performed toward her should increase her gratitude.

Uncertainty derives from two conceptually distinct sources: either the expected behaviors are not fulfilled as expected (the landlord turns philanthropist); or other unexpected elements intervene and modify the conditions of the behavior (a new law regarding tenants is issued). But an action, if engaged, corresponds to a probability estimation of the different outcomes.

To sum up, an ordinary action rests on intentions. In addition, it also depends upon:

1. An anticipation of the normal consequences of given conduct that causally lead to the implementation of these intentions; these consequences presuppose causal chains that bring about the expected results. These causal chains can be founded upon institutional regularity, where people are expected to ordinarily behave in certain ways.
2. An analytical focus on the relevant causal properties of the action which is to be carried out; other important elements, while less relevant but necessary to the achievement of the action, are left aside.
3. A tendency toward the abstraction of elements that could causally impede the realization of this expected regularity, arising from a belief that such elements will not intervene.
4. An estimation of the probability associated with the various elements that intervene in the course of action.

Social science explanation rests upon regularities in causal action

As stated above, explanation in the social sciences is based on the causal features of ordinary actions, and corresponds to these features. If this explanation is linked to a prediction, it has the same general features noted previously; if it corresponds to past events, it has to select the relevant elements that led to a specific outcome. This is possible only on the basis of the regularities in actions. For instance, in his blog Gary Becker warns against current plans for the limitation of high levels of remuneration:

> There is no good reason, however, for the government to interfere and impose limits on salaries and severance pay. Controls over wages and salaries have never worked well, and only encourage myriad ways to get around them, including generous housing allowances, vacation homes, easy access to private planes, large pensions, and other fringe benefits. There develops a war between the government's closing of loopholes, and the ingenuity of accountants and lawyers in finding new ones. (Becker 2008)

The rationale is that given the assumed motivations of the managers, and if a shift occurs in public policy regarding golden parachutes, it is anticipated that the new rules will lead to specific consequences (which Becker dislikes). The important point for our purposes is that the whole argument rests on:

1. the description of typical regularities in behavior, depending on typical motives, and of typical consequences of those motives;
2. analytical focus on the process, regardless of all the other elements (the existence for instance of corporations, of a market for corporate

executives, a legal system and so on, with all the corresponding actions);
3. the abstract exclusion of all elements that would interfere with the expected outcome: the fact that executives could suddenly wish to lead a modest life; or that new rules would prevent them from obtaining generous housing allowances;
4. an estimation of the probability of the outcomes: here Becker considers that government action would involve a false assessment of the probability of envisioned consequences.

This explanatory device is exactly the same as that theorized by J.S. Mill (1843). It is also similar to the way in which Macy *et al.* describe abstractness in this volume (Chapter 12).

What is then exactly "understood" when an explanation of such actions is given (which leads moreover, in Becker's example, to a prediction)?

First, the fact that a given action is intentional does not mean that it is "understandable" as such for an observer (although Dilthey initially, and Collingwood (1946) and Dray ([1964] 1993) thereafter have defended that kind of view); or that it provides a firm basis for explanation. It depends on what is meant by "understanding." Understanding an intention can mean two different things. The first is the notion of simple recognition of the existence of an intention in someone. But, more ambitiously, understanding an intention is in some way the recognition of the "relevance" of such an intention in particular circumstances, given the actor's situation. Therefore, referral to the intentional aspect on an action is not sufficient guarantee of understanding it in this more ambitious meaning of the word. For instance, when we learn that the Nazis endeavored to exterminate all Jews, this action is clearly an intentional one. We understand the intention, if understanding means the recognition of the existence of the intention. But this does not make it understandable in the second sense: why was such an intention formed? Intentional actions that we observe in others often seem to us absurd, foolish and irrelevant. We would certainly not do the same were we in their shoes, and we do not find the action relevant even in any weak sense. To put it another way, the fact that others follow rules (Nazis are supposed to kill Jews) does not render the intention any more understandable for us.

To provide an explanation of an action we need to refer it to other elements that provide a basis to the explanation. These elements typically are:

1. either another intention, or a cause that produces the intention to be explained;

2. a link between the other intention, or the cause, and the action to be explained; the link existing because there is a regularity which constitutes the link. The regularity is discovered through experience, and depends on its features.

In the simple case of instrumental action, where means are chosen in order to achieve an end, the explanation rests both on the link that exists between the means and the end, and on the fact that an intention to reach an end implies (cognitive constraint) the selection (practical constraint) of the means. There might also be institutional constraints (some means are socially accepted, others are not). Someone can give up trying to attain an end because he does not accept the necessary means to achieve it, but this rests on the fact that there is a practical regular relation between the means and the end. These causal relationships are discovered through our experience, and correspond to very different situations. Explanation of an instrumental action rests on two obvious dimensions, which are elements not considered to be explained of themselves: the link between an end and the means to achieve it, and the link between an intention and the choice of the means. Those two dimensions are thought so obvious as *to need no explanation*; they are the basis on which the explanation is built. For instance, when I take the train to Marseille: I can explain the fact that I take the train by the fact that I have decided to go to Marseille, and that taking this train allows me to go to Marseille, and that I have no objection toward taking the train. The explanation does not rely on intentional action as such (the intention of going to Marseille is not explained) which is not a sufficient basis for a genuine explanation; nor does it rest on our ability, in general, to understand others; there are many things we do not understand. The explanation rests here on two dimensions which are considered to be evident, and which are two causal regularities:

1. The empirical regularity that taking a specific train allows someone, under normal circumstances, to reach Marseille (practical constraint; it should be noted here that many institutions are here involved, in particular train companies).
2. The regular fact that the decision to go to Marseille can lead causally to the intention of choosing (*ceteris paribus*) the available means to go there (taking the train): cognitive constraint, considering that the means do not arouse any kind of objection (they are accepted).

In other words, I wish to stress that the explanation necessarily rests on the description of typical regular causal relations, elements *that are not*

explained as such: on the one hand the fact that taking a train allows us to go to Marseille; and given the fact that she has an intention, a person selects the acceptable means to achieve her intention.

When we turn to analyze an action based on values we encounter exactly the same structure of explanation. If we want to explain, for instance, why some people do not have abortions by referring to the fact that they are Catholics, we base our explanation on two given regularities: the fact that Catholic values, in the current organization of Catholicism, imply a refusal of abortion, and the fact that some women do not have abortions so that they might be good Catholics. The explanation rests on descriptive regularities, and involves elements that are not explained as such, but which are necessary to the explanation: the fact that Catholic values currently imply a refusal to have an abortion; and the fact that many Catholics take into account, regarding abortion, the consequences for their faith. Catholicism as such is not explained, nor the intention of being a Catholic.

So any explanation of action rests on regular features that establish a link between different aspects of actions; these links are not themselves explained. They are the building blocks of the explanation; I have intentionally left aside all debate over causes and reasons which make things more difficult. But that does not alter the main argument (see, for instance, Skorupski 2001).

Similar things can be said of explanations concerning natural phenomena. To explain why the planting of legumes increased the fertility of croplands we have to refer it to a causal regularity, which is that plants do not all use the same nutrients, and do not introduce the same elements into the soil. A causal regularity of this kind forming the basis of the explanation is not explained as such. It is a regularity which is given, and on which the explanation rests. There might be further attempts to explain these regularities by referring them to "finer grain" regularities (Elster 2007). Harré acknowledges that "mechanistic" regularities rely ultimately on other lower-level regularities (see the Introduction to this volume). There might for instance be attempts to explain intentional action as such (Gibbard 1990); in the social sciences a convenient endpoint in this regression would be a sound theory of common-sense, if it exists (Abbott 2004; Boudon 2006). Sperber (Chapter 3, this volume) argues that we should not stop at this level.

Hence any process of explanation relies on causal regularities that are discovered; as regards social action, it is not the intentional character of an action as such that enables us to understand it, but regular connections between intentions and practical and institutional constraints which allow certain consequences to be reached when certain actions

are engaged, based on typically triggered intentions. Any explanation depends on such regularities. The regularities are sometimes explained by other regularities, sometimes not. For instance, it is possible to explain many different types of behavior by referring to a notion of interest. Nevertheless, a fundamental issue in social science remains explaining why people do not have the same choices and preferences, that is, the same interests.

When it comes to opportunities, it is clear that different opportunities causally create the *possibility* of different actions. For instance, the development of agriculture in Britain between the eighteenth and nineteenth centuries created the *possibility* of an increase in population; the fact that agricultural development increased the global number of population (since a larger number of people could be fed) was linked to the fact that a smaller percentage of the population was necessary for agricultural production; therefore those no longer employed in agriculture could find employment in newly expanding industry. We therefore here have typical opportunities that "trigger" new possibilities of action (and consequently actions themselves, although they were not necessary). Regarding beliefs, it is similarly clear that different levels of information can prompt different beliefs, based on different perceptions of evidence. The explanation here rests on the fact that different information causally produces different beliefs. What is hard to explain is why people who are in similar situations do not have the same desires and intentions.

Is a Hempelian approach to explanation non-mechanistic?

Any explanation is based on causal regularities; those regularities, regarding individual actions, are of three kinds:

1. cognitive or psychological constraints (I would include here emotions);
2. pragmatic constraints, which refer at the same time to the way a definite action can be achieved, and to the environmental limitations which allow certain actions to be realized and render others impossible;
3. institutional constraints (social rules which organize social life and intervene in the realm of pragmatic constraints).

Since any explanation of an action relies on such causal regularities, should we distinguish mechanism-based explanations and

covering-law explanations, since they both necessarily rely on and refer to regularities?

I will first mention the very important fact that Hempel (1948) himself, when he describes social sciences explanations, chooses examples which can clearly be interpreted in terms of "mechanisms": when he describes a panic on the stock exchange he relies at the same time on stable given motivations and on predictable outcomes based on these motivations, within a definite institutional framework. Since, moreover, he introduces the necessity of a probabilistic approach, it is clear that his analysis can be appropriately interpreted in terms of a mechanistic explanation. I will quote Hempel at length so as to make his reasoning clear:

> Let us now consider an illustration involving sociological and economic factors. In the fall of 1946, there occurred at the cotton exchanges of the United States a price drop which was so severe that the exchanges in New York, New Orleans, and Chicago had to suspend their activities temporarily. In an attempt to explain this occurrence, newspapers traced it back to a large-scale speculator in New Orleans who had feared his holdings were too large and had therefore begun to liquidate his stocks; smaller speculators had then followed his examples in a panic and had thus touched off the critical decline. Without attempting to assess the merits of the argument, let us note that the explanation here suggested again involves statements about antecedent conditions and the assumption of general regularities. The former include the facts that the first speculator had large stocks of cotton, that there were smaller speculators with considerable holdings, that there existed the institution of the cotton exchanges with their specific mode of operation, etc. The general regularities referred to are – as often in semi-popular explanations – not explicitly mentioned; but there is obviously implied some form of the law of supply and demand to account for the drop in cotton prices in terms of the greatly increased supply under conditions of practically unchanged demand; besides, reliance is necessary on certain regularities in the behavior of individuals who are trying to preserve or improve their economic position. Such laws cannot be formulated at present with satisfactory precision and generality, and therefore, the suggested explanation is surely incomplete, but its intention is unmistakably to account for the phenomenon by integrating it into a general pattern of economic and socio-psychological regularities. (Hempel 1948, 1965: 251–2)

It is clear from this quotation that in Hempel's description covering-law explanations naturally include mechanism-based explanations, since the latter are themselves based on causal regularities: it is important moreover to note that in Hempel's argument, institutions and their regularity are mentioned and belong to the story. The whole possibility of the explanation rests on:

1. motivational regularities;
2. practical regularities;
3. institutional regularities.

I believe that all these regularities do not deserve to be treated as "laws," nor as "quasi-laws." I will come to this point below, but these regularities are part of the rationale of determining the possibility of an explanation in social science.

Contrary to this line of argument, Hedström, following Harré (1972), expresses deep reservations about the similarity between mechanism-based explanations and covering-law explanations. I will not give all the details of his discussion, which he summarizes in this way:

> Although the covering-law model has many attractive features, I do not think that the model as such is particularly useful for sociology. The main reasons are the following:
>
> 1. The deductive nomological model is not applicable because the deterministic social laws that it presupposes do not exist.
> 2. The inductive–probabilistic model is not useful as an explanatory model because (a) it allows for and thereby legitimizes superficial theories and explanations, and (b) it does not give action and intentional explanations the privileged role they should have. (Hedström 2005: 20)

The third point (2b), the fact that covering-law explanations do not give action and intentions the role they deserve is surprising, since all the examples given by Hempel rest on those elements. If we consider that explanations have to rely on individual actions, the regularities are to be found at the level of those actions and of their typical consequences. The important point I have made is that actions do proceed on the basis of causal regularities; therefore explanations depend on the same regularities.

Concerning the second point (2a), I do not consider Salmon's objection (Salmon 1971) to be definitive. It rests in my view on an inappropriate and restrictive interpretation of Hempel's theorization. Let us describe Salmon's objection in Hedström's terms:

> There exist statements that fulfil all of Hempel's logical requirements but which nevertheless are not explanatory. The following explanation is a case in point. If we wanted to explain the fact that Peter did not become pregnant, the following line of reasoning would appear to be acceptable from Hempel's perspective (adopted from Salmon 1971)
>
> No one who regularly takes birth-control pills becomes pregnant
> Peter regularly takes birth-control pills

(therefore) Peter did not become pregnant

The fact to be explained can be logically deduced from the premises – both of which can be assumed to be true – but the explanation is nevertheless incorrect because it refers to the wrong causal mechanism. (Hedström 2005: 16)

It seems to me that this presentation of the structure of an explanation does not correspond to the one given by Hempel, which is not just a syllogism, since the laws which are used in the explanation must relate accurately and appropriately to the phenomenon to be explained. Let us refer to the way Hempel himself theorizes the deduction:

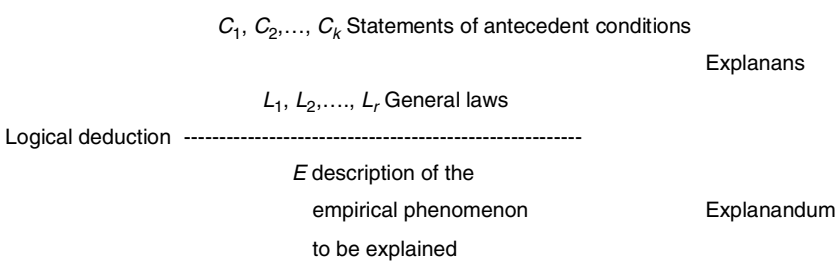


Figure 9.1 Hempel's explanation schema (from Hempel 1965: 249).

The important thing that I wish to stress is that to reach a satisfactory deduction, Hempel mentions a *plurality* of relevant laws that pertain to the phenomenon to be explained, and are empirically relevant. Therefore it seems difficult to construct a deduction around only one law if we are to accurately describe and explain a given phenomenon. In the case of the birth-control pills, one general law which is currently very certain is that men cannot have children; a consequence of this is that birth-control pills cannot work on men. Therefore, among the relevant general laws to be taken into account in reasoning about contraceptive pills is that they cannot work on men, since there is another law stating that men cannot have children. Salmon's reasoning would be therefore incomplete regarding the alleged deduction. The relevant laws have to be selected in accordance with the phenomenon to be studied; otherwise the whole process is meaningless. The explanation is not only a formal logical deduction; it is based on given knowledge about empirical phenomena.

Finally, regarding the first point advocated and which seems to me the most important, there appears to be no reason to exclude from action field deterministic laws unilaterally and deny their existence. Since action is always based on causal regularities, these regularities have moreover consequences at the outcome level that are themselves regular (see Bunge 2004 and Opp 2005).

Hedström's discussion begins in fact from the realm of natural sciences, where he seems to rule out covering-law explanations which he considers to be unsatisfactory, to the benefit of mechanism-based explanations. Let us follow the strychnine example he takes:

> It would be possible to statistically estimate the parameters of an equation describing the relationship between the intake of, say, strychnine and the risk of dying.
>
> […]
>
> Such an explanation seems wanting however. When posing such questions in a scientific context we normally expect answers that not only state *that* the event was likely because this is what happened in the past, we also want to know *why* this is so.
>
> […]
>
> By pointing to how strychnine typically inhibits the respiratory centre of the brain and to the biochemical processes typically responsible for such paralysis, we provide a mechanism that allows us not only to predict *what* is likely to happen but also to explain *why* (Bunge, 1967). For these reasons, I am inclined to agree with von Wright that it is better "not to say that the inductive-probabilistic model [of Hempel] explains what happens, but to say only that it justifies certain expectations and predictions (von Wright 1971: 14)." (Hedström 2005: 17)

In fact, introducing such a mechanism-based explanation does not introduce any novelty regarding the problem of covering-law explanation: since the explanation of the manner in which strychnine inhibits the respiratory center of the brain itself rests inescapably on a covering-law explanation, as well as the explanation of the biochemical processes typically responsible for such paralysis. These are regularities which are not explained as such, but are described; but they allow us to explain given empirical phenomena. We follow in this example a path where we proceed from more particular descriptions to more general ones upon which the particular can be based. But the more general regularities are regularities that are not explained as such.

As regards law in the field of action, as I have argued above any explanation of human action rests on psychological-cognitive, practical and institutional regularities. Without such regularities, we would have no basis for explanations. There are, therefore, no reasons to rule them

out. For instance, when we say that the British agricultural revolution produced an increase in population and thereby facilitated the industrial revolution, this reasoning rests on two "regularities" that can be easily accepted (from Polanyi 1944):

1. the increase in agricultural productivity makes possible an increase of population;
2. when food can be supplied for a larger population without having the whole population dedicated to agricultural activities, some other activities can be developed.

Therefore the increase of population added to the fact that the enclosure movement had the practical consequence of eliminating many people from agricultural activities, liberating a large workforce for the developing industry. All these elements are true, and they are based on practical regularities. We should not therefore consider that there are no laws regarding social actions. The more difficult question is to determine whether all causal regularities are laws. I will not address this question in detail in this chapter, but make a few concluding (but very important in my view) remarks.

Some regularities can clearly be considered to be laws: the more food that is available, the larger a population can be (even though it is not necessary that the population does in fact increase its size; this depends on other factors).

Second, many regularities are not strictly deterministic laws, since it is always possible for a singular individual to fail to behave according to the predicted regularity. Nevertheless, it is a regularity, since people in certain circumstances do tend to behave in certain ways, even if they do not do so all the time. This can be understood as a probabilistic law, in the manner described by Hempel.

Third, many of these regularities depend on the existence of institutions that could be organized differently; not as in nature, where it would be absurd for someone to decide to change its laws. This is a major and significant difference. Rules can often be changed and, when changed, they produce new regularities. But when established they allow us to predict major regularities.

It is with this essential point that I wish to conclude. In our mechanism-based explanations we frequently rely on institutions, norms and rules that trigger regular behaviors and social outcomes whenever those rules exist and are accepted. And so the explanation relies on such causal regularities, on the basis of the existence of the rules and their general acceptance causally determining social outcomes. But

those rules could be different. That is why those rules are not laws, nor "quasi-laws" in the natural sciences meaning, although we need them in order to produce causal explanations in the social world. A murderer can "causally" be led to death, through a series of steps, in the United States. Would the death penalty be abolished, he would no longer be causally sentenced to death. He might causally be confined to prison for his entire life. Therefore, it should be relevant to mention causal regularities as a basis for causal explanations in the social world, but not to consider all of them to depend on laws of the type to be found in the natural world.

REFERENCES

Abbott, Andrew. 2001. *Time Matters. On Theory and Method*. University of Chicago Press.

2004. *Methods of Discovery. Heuristics for the Social Sciences*. New York and London: W.W. Norton and Company.

Becker, Gary. 2008. *Government Equity in Private Companies: A Bad Idea*. Becker–Posner Blog October 5.

Boudon, Raymond. 2006. *Renouveler la démocratie. Eloge du sens commun*. Paris: Odile Jacob.

Bunge, Mario. 1967. *Scientific Research*. Berlin and New York: Springer-Verlag.

2004. "How does it work? The search for explanatory mechanisms," *Philosophy of the Social Sciences* 34(2): 182–210.

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

Collingwood, Robin George (1946, revised edition 1993). *The Idea of History*. Oxford University Press.

Demeulenaere, Pierre. 2008. "La définition des coûts et avantages, entre identité et diversité des préférences," *Revue européenne des sciences sociales XLVI*(140): 175–95.

Demsetz, Harold. 1967. "Toward a theory of property rights," *American Economic Review* 57: 347–59.

Descombes, Vincent. 1996. *Les institutions du sens*. Paris: Minuit.

Dilthey, Wilhem. 1883. *Introduction to the Human Sciences*. Detroit: Wayne State University Press..

Dray, William H. [1964] 1993. *Philosophy of History*. Upper Saddle River: Prentice Hall.

Eggertsson, Thrainn. 1990. *Economic Behavior and Institutions*. Cambridge University Press.

Elster, Jon. 1989. *Nuts and Bolts for the Social Sciences*. Cambridge University Press.

2007. *Explaining Social Behaviour. More Nuts and Bolts for the Social Sciences*. Cambridge University Press.

Geertz, Clifford. 1973. *The Interpretation of Cultures*. New York: Basic Books.
Gibbard, Allan. 1990. *Wise Choices, Apt Feelings*. Cambridge, MA: Harvard University Press.
Harré, Rom. 1972. *The Philosophies of Science. An Introductory Survey*. Oxford University Press.
Hedström, Peter. 2005. *Dissecting the Social. On the Principles of Analytical Sociology*. Cambridge University Press.
Hempel, Carl (with Paul Oppenheim) 1948. "Studies in the logic of explanation," *Philosophy of Science* 15: 135–75.
Hempel, Carl G. 1965. *Aspects of Scientific Explanation and other Essays in the Philosophy of Science*. New York and London: The Free Press.
Hollis, Martin. 1994. *The Philosophy of Social Science. An Introduction*. Cambridge University Press.
Hume, David. 1758. *Enquiries concerning Human Understanding and the Principles of Morals*, edited by L.A. Selby-Bigge, revised by P.H. Nidditch. Oxford: Clarendon Press, 1975. London: Routledge and Kegan Paul.
Mantzavinos, Chrysostomos. 2001. *Individuals, Institutions, and Markets*. Cambridge University Press.
Menger, Carl. [1883] 1996. *Investigations into the Method of the Social Sciences*. Grove City: Libertarian Press.
Mill, John Stuart. 1843. *System of Logic, Ratiocinative and Inductive*, in *The Collected Works of John Stuart Mill*, vols. VII and VIII. London and Toronto: Routledge and University of Toronto Press, 1973.
Opp, Karl-Dieter. 2005. "Explanations by mechanisms in the social sciences. Problems, advantages and alternatives," *Mind and Society* 4: 163–78.
Passeron, Jean-Claude. 2006. *Le raisonnement sociologique*. Paris: Albin Michel.
Polanyi, Karl. 1944. *The Great Transformation*. Boston: Beacon Press.
Rickert, Heinrich. 1902. *Die Grenzen des naturwissenschaftlichen Begriffsbildung*. Tübingen: Mohr.
Salmon, Wesley C. 1971. *Statistical Explanation and Statistical Relevance*. University of Pittsburgh Press.
Sen, Amartya K. 1977. "Rational fools: a critique of the behavioral foundations of economic theory," *Philosophy and Public Affairs* 6(4): 317–44.
Shackle, George L.S. 1972. *Epistemics and Economics. A Critique of Economic Doctrines*. Cambridge University Press.
Simmel, Georg. 1923. *Die Probleme der Geschichtsphilosophie. Eine erkenntnistheoretische Studie*. Munich and Leipzig: Von Duncker et Humblot.
Skorupski, John. 2001. "Rationality – instrumental and other," in R. Boudon, P. Demeulenaere and R. Viale (eds.), *L'explication des normes sociales*. Paris: P.U.F.
Veyne, Paul. 2008. *Foucault. Sa pensée, sa personne*. Paris: Albin Michel.
Von Wright, G.H. 1971. *Explanation and Understanding*. Ithaca: Cornell University Press.

Weber, Max. [1904] 1991. "Die 'Objektivität' sozialwissenschaftlicher und sozialpolitischer Erkenntnis," in *Schriften zur Wissenschaftslehre*. Stuttgart: Reclam.

Winch, Peter. 1958. *The Idea of a Social Science and its Relation to Philosophy*. London: Routledge.

Windelband, Wilhelm. [1894] 2000. "Histoire et science de la nature (Discours de Rectorat)," *Les études philosophiques* 1: 1–16.

Part III

Approaches to mechanisms

10 Youth unemployment: a self-reinforcing process?

Yvonne Åberg and Peter Hedström

Introduction

During the past two decades, social scientists and policy-makers, particularly in the United States, have paid increasing attention to neighborhood-based social interactions. Much of this upsurge in interest can be traced to the writings of William Julius Wilson on the importance of neighborhood characteristics in explaining inner-city social problems in the United States (e.g. Wilson 1987; see also Sampson *et al*. 2002).

This research is closely aligned with core themes in sociological theory and research, because social interactions are at the heart of sociology.[1] A focus on social interactions is particularly salient among analytical sociologists because analytical sociology explains by detailing mechanisms through which social facts are brought about, and these mechanisms invariably refer to individuals' actions and the relations that link them to one another (see Hedström and Bearman 2009).

This chapter is concerned with the role of peer-based social interactions in explaining the length of unemployment spells and spatial variations in unemployment levels. We seek to specify in some detail why social interactions are important, and we use unique population-level panel data to assess their importance.

The questions addressed in the chapter are highly specific but the mechanisms we focus upon and the methods we use are much more general. The chapter is directly related to a growing body of empirical and theoretical work on the importance of social interactions for

We wish to thank Mary Brinton, Magnus Brygen, Christopher Edling, John Goldthorpe, Carl le Grand, Ann-Sofie Kolm, Thomas Korpi, Charles Manski, Peter Marsden, Michael Sobel, Arthur Stinchcombe, and Michael Talhlin for their valuable comments on previous versions of the chapter. The research reported on here has been made possible by grants from the Office of Labour Market Policy Evaluation, the Bank of Sweden Tercentenary Foundation, and the Swedish Council for Working Life and Social Research.

[1] As Max Weber once expressed it: "Sociology … is a science concerning itself with the interpretative understanding of social action and thereby with a causal explanation of its course and consequences" (Weber [1921–2] 1978: 4).

various social and economic processes, and it is a direct continuation of our own previous research on the role of social interactions in explaining suicides (Hedström *et al.* 2008), and the role of social interactions in explaining couples' decisions to divorce (Åberg 2009).

The chapter is organized as follows. In the next section, we define more precisely what we mean by a social-interaction effect, and we distinguish between different types of social-interaction effects on the basis of how the action of one individual influences that of another. In the third section, we develop a theoretical model that allows us to examine how unemployment levels are likely to be affected if social-interaction processes are at work. In the fourth section, we use unique population-level panel data to examine the importance of social interactions for explaining youth unemployment in the Stockholm metropolitan area during the 1990s. First, we focus on variations in unemployment levels between different neighborhoods, and then we examine whether the unemployment level in an individual's peer group systematically influences the time it takes for the individual to leave unemployment. Finally in the fifth section, we summarize our results and discuss some wider implications of our findings.

Social interactions

Before discussing why social interactions are likely to be important in this context, we must define more precisely what we mean by a social-interaction effect, and how social-interaction effects differ from other related types of behavioral uniformities.

What is a social-interaction effect?

One can distinguish between at least three types of process that can result in individuals in a group acting in a similar manner, and only one of these has anything to do with social interactions. We can use the following example from Max Weber to clarify the differences between them:

> Social action is not identical with the similar actions of many persons … Thus, if at the beginning of a shower a number of people on the street put up their umbrellas at the same time, this would not ordinarily be a case of action mutually oriented to that of each other, but rather of all reacting in the same way to the like need of protection from the rain. (Weber [1921–2] 1978: 23)

This piece of everyday behavior is not "social action" explained by some form of interaction between the people on the street, but is due to an *environmental effect*, in this case a rainfall that made all individuals adjust their behavior in a similar manner. Such environmental effects can easily be mistaken for interaction effects. Assume that Weber's

rainfall started at one end and gradually spread along the street. The pattern of umbrella use would then "diffuse" in a way that could easily give the impression of it being the result of a genuine social-interaction effect, where one individual's umbrella use increased the likelihood that adjacent individuals would use their umbrellas.

Even if during said rainfall we observed that the frequency of umbrella use was higher among those walking on one street than on another, this would not necessarily mean that we were observing the outcome of some sort of interaction process. It could simply be due to a *selection effect*, in this case that individuals with a preference for using umbrellas for some reason ended up walking on one of the streets rather than on the other. For example, if the stores on one street catered to young people, the observed pattern simply could be due to an age-based selection effect since young people are less likely to use umbrellas than older people are. If we do not take such differences into account we may easily mistake selection effects for social-interaction effects.

Environmental effects and selection effects differ from social-interaction effects in that the correlated behavior they give rise to has nothing to do with individuals influencing one another. A *social-interaction effect* exists if and only if it was the umbrella use of others that influenced the focal individual's use of the umbrella. For example, some individuals may hesitate to use their umbrellas because using an umbrella could indicate to others that they were excessively concerned with their appearance. Although they would have liked to use their umbrellas, they decided against it in order not to appear excessively vain. But once others started to use their umbrellas they quickly followed suit. This would then be an example of a social-interaction effect, because it was the actions of others that influenced their decisions whether to use an umbrella or not.

The distinctions introduced so far are summarized in the upper part of Figure 10.1 and may be expressed as follows: an environmental effect is operative if we do what we do because we are where we are. A selection effect is operative if we do what we do because we are who we are. And finally, a social-interaction effect is operative if we do what we do because others do what they do.

Types of social-interaction effects

Social-interaction effects can arise for rather different reasons, and in order to better understand why we observe what we observe it is useful to try to distinguish between them. As suggested in Hedström (2005), one can distinguish between at least three broad types of social interaction: opportunity-based, belief-based and desire-based interactions

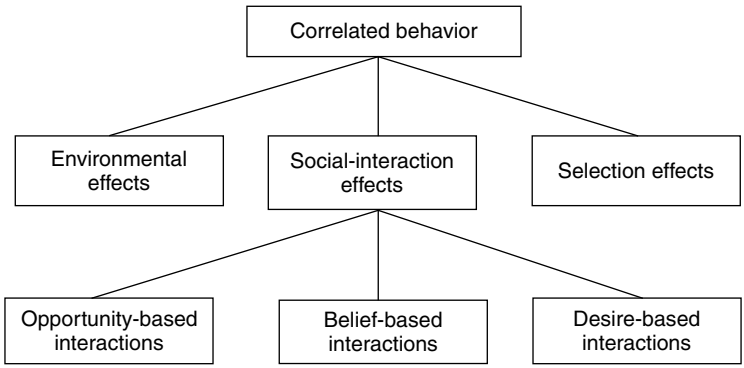


Figure 10.1 Sources of correlated behavior among individuals.

(see also Manski 2000 for similar distinctions). Consider the case of an unemployed individual and an action that influences the likelihood that the individual remains unemployed. How can the unemployment level among others influence this action? The general answer is that this can occur in three different ways:

1. the unemployment level among others can influence the focal individual's *opportunities*, and thereby his or her choice of action;
2. it can influence the focal individual's *beliefs*, and thereby his or her choice of action; and
3. it can influence the focal individual's *desires*, and thereby his or her choice of action.

Let us briefly elaborate on each of these three types of social interaction.

As observed by Granovetter (1995) and others, many individuals obtain their jobs via informal social contacts with friends and acquaintances. Friends and acquaintances pass on information about jobs to prospective job candidates, and information about potential job candidates to employers. If the unemployment level is high among friends and acquaintances, information about vacant jobs will not reach the focal individual to the same extent as if friends and acquaintances were employed. Therefore the focal individual's probability of finding out about vacant jobs will be negatively influenced by the unemployment level among others. This is an example of an opportunity-based interaction effect.

The individual's likelihood of leaving the unemployed state is also likely to be influenced by the individual's beliefs about the jobs that he or she

can expect to get. Traditional decision and search theory suggests that those who expect to get a job, particularly a high paying job, are willing to invest more time and energy into job search than those with bleaker prospects. To the extent that these beliefs are partly influenced by the experiences of others, we have an example of a belief-based interaction effect. One example of belief-based interaction is the so-called "discouraged worker effect," i.e. the notion that a high unemployment level may discourage individuals from looking for work because they do not expect to find any (e.g. Schweitzer and Smith 1974). Another example of belief-based interaction arises when other individuals serve as role models. One reason for Wilson's concern about the exodus of middle-class families from many ghetto neighborhoods in the United States, for example, was precisely such belief-based interaction effects: "the very presence of these families … provides mainstream role models that help keep alive the perception that education is meaningful, that steady employment is a viable alternative to welfare, and that family stability is the norm, not the exception" (Wilson 1987: 56). In both the discouraged-worker and the role-model case, unemployment among others influences an individual's beliefs and subsequent actions in such a way that his or her chances of leaving the unemployed state are altered.

One reason for expecting desire-based interactions to be important in the context of unemployment is the existence of strong normative pressures to earn one's living. Being unemployed usually means that one cannot live up to this norm, and this may bring about feelings of shame or embarrassment (Elster 1989). In Zawadski and Lazarsfeld's classic study of the psychological effects of unemployment one can find the following autobiographical note of an unemployed mason:

> How hard and humiliating it is to bear the name of an unemployed man. When I go out, I cast down my eyes because I feel myself wholly inferior. When I go along the street, it seems to me that I can't be compared with an average citizen, that everybody is pointing at me with his finger. I instinctively avoid meeting anyone. Former acquaintances and friends of better times are no longer so cordial. They greet me indifferently when we meet. They no longer offer me a cigarette and their eyes seem to say, "You are not worth it, you don't work." (Zawadski and Lazarsfeld 1935: 239)

Although the details of the Polish mason's experiences may seem a bit dated, they highlight an important aspect of the unemployment experience that is as relevant today as it was in Poland in the 1930s: being unemployed is often associated with strong feelings of shame and embarrassment.

Desire-based interactions are also likely to be important for reasons that are unrelated to social norms. Being the only unemployed

individual, for example, is likely to be a rather lonely and dull existence compared to one in which many of one's friends and acquaintances are also unemployed. When friends are unemployed and available for company, daily activities are more stimulating then when none of one's friends have the time to socialize during daytime.[2] Thus, an increase in unemployment among an individual's friends and acquaintances is likely to reduce the social and psychological costs of being unemployed through several different types of mechanism (see also Clark 2001).

Why are social interactions likely to be important for aggregate unemployment?

The analysis of social interactions is fraught with difficulties. Not only is it difficult to establish their very existence, it is also difficult to know exactly why we observe the interaction effects we observe. Ignoring them is not a viable option, however, since in many cases, they are likely to be of considerable importance. When they operate, endogenous processes are likely to be important for changes in aggregate unemployment. A defining characteristic of an endogenous process is that the number of individuals who act in a certain way at a certain point in time in itself partly explains how many others will adapt the behavior at a later point in time. Variations in the overall unemployment level probably have little to do with such processes, but in some geographical areas or peer groups, unemployment levels may be greatly influenced by such processes.

In order to get a better idea of how aggregate unemployment is likely to be affected by social interactions, we use a simple differential equation model. For illustrative purposes, we focus exclusively on desire-based interactions. The effects of opportunity-based and belief-based interactions are likely to operate in the same direction, however, and therefore they are likely to amplify rather than counteract the effects revealed below.

We assume that individuals can be in one of two states, they can be unemployed or they can be employed, and we consider what may happen if the unemployment level among others influences a focal

[2] As part of this study we conducted a series of in-depth interviews with unemployed individuals in the Stockholm region. In these interviews the importance of the unemployment of friends and acquaintances is a recurrent theme. One person expressed himself in the following way: "Now with this beautiful weather it is wonderful to be unemployed. I have many friends who are unemployed, so I can meet them during the days. Instead of being locked up inside an office all the day one can be outside and play soccer … But if all my friends were working I would want to do so as well. Otherwise I would just sit at home without anything to do."

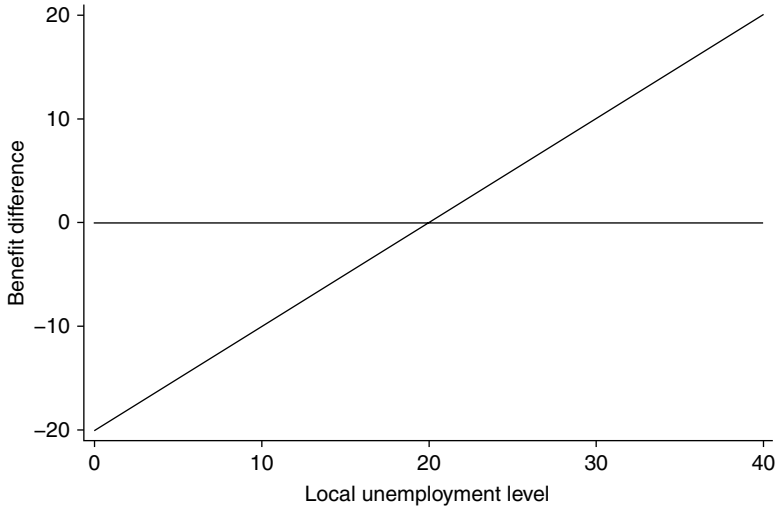


Figure 10.2 Hypothetical benefit difference between being unemployed and employed. "Benefit difference" is equal to the social, economic, and psychological benefits of being unemployed minus the social, economic and psychological benefits of being employed.

individual's unemployment experience and behavior. Figure 10.2 illustrates what we have in mind.[3]

When the unemployment level is low, it is much more advantageous to be employed than to be unemployed. But when the unemployment level increases, the difference is reduced. In this hypothetical example the combined social, economic and psychological benefits of being unemployed exceed those of being employed once the unemployment level in the relevant peer group exceeds 20 percent.

The benefit function in Figure 10.2 illustrates a so-called "tipping point," which in this case occurs when 20 percent are unemployed. The tipping point is an unstable equilibrium, and a slight exogenous "push" in any direction can set in motion an endogenous process with considerable long-term consequences. If external events make the local unemployment level exceed the tipping point, the benefit difference becomes positive, which will further increase the unemployment level. If the unemployment level falls below the tipping point, a parallel

[3] This part of the model has been greatly inspired by Schelling's (1978) work on discrete choice processes with social externalities.

endogenous process (operating in the opposite direction) leads to a reduction in the unemployment level.

It seems reasonable to assume that the size of this benefit difference influences how *rapidly* individuals move to the most advantageous alternative. For example, when the benefit of being unemployed falls relative to the benefit of being employed, individuals can be expected to invest more time and effort in searching for jobs, which is likely to shorten their unemployment spells. The actual change in the number of unemployed obviously also depends upon how many individuals are susceptible to change. In the extreme case, when there are no unemployed individuals, the unemployment level cannot be reduced further no matter how negative the benefit difference is. Thus, everything else being the same, the change in the number of unemployed individuals per unit of time brought about by the endogenous process is likely to be influenced by:

1. the benefit difference between being unemployed and employed; and
2. the number of employed and unemployed individuals at that point in time.

As emphasized above, much of the variation in unemployment that we observe is likely to have little to do with such endogenous processes. Labor market conditions change, and this changes the unemployment level irrespective of the beliefs and desires of the individuals. In the model presented below, we capture exogenous causes of unemployment with a "business cycle effect." Since we are not interested in this effect as such, we may as well model it in the simplest possible fashion, and we therefore represent it by a sine function. Given these assumptions, the change in the local unemployment level can be analyzed with the following ordinary differential equation:[4]

$$\frac{du}{dt} = \begin{cases} (1-w)\times(-+\beta\times u_t)\times u_t + w\times\sin(r\times t) & \text{if} \quad v_t \leq 0 \\ (1-w)\times(-+\beta\times u_t)\times(N-u_t) + w\times\sin(r\times t) & \text{if} \quad v_t > 0 \end{cases} \tag{1}$$

where

$\frac{du}{dt}$ = change in the number of unemployed per unit of time,

u_t = number of unemployed at time t,

[4] The model is a slightly modified version of a model developed in Åberg (2000).

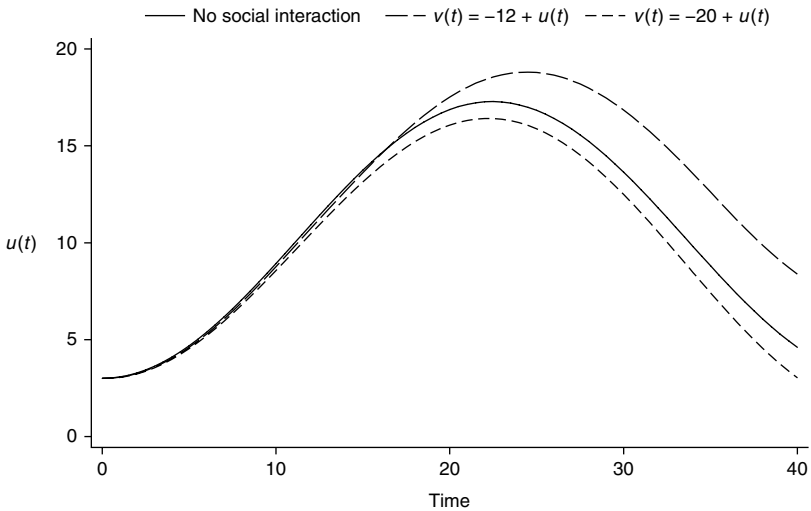


Figure 10.3 The influence of social interactions on the local unemployment level.

$v_t = \alpha + \beta \times u_t$ = the benefit difference between being unemployed and employed at time t, assumed to be a linear function of the number of unemployed at time t,

N = the total number of individuals,

w = a weight that dictates how important the exogenous business cycle component is relative to the endogenous component, and

r = a parameter that dictates the amplitude of the business cycle effect.

By solving Equation (1) numerically, we are able to tell how the unemployment level develops over time, given different assumptions about the benefit functions; that is, the solution will yield the unemployment level as a function of time for a given benefit function.

Figure 10.3 shows how the unemployment level (in a particular peer group or neighborhood) would change over time given these assumptions.[5] The solid line shows how the unemployment level would change if there were no social-interaction effects, and the dashed lines show how the unemployment level would evolve with such effects. As the top dashed line shows, the endogenous process leads to a substantial

[5] The w-value is set to 0.9995, the r-value to 0.14, and the initial unemployment level to 3. These values were chosen to generate a realistic variation in the unemployment levels.

increase in the unemployment level once it passes the tipping point (which in this example is 12). The lower dashed line shows what happens when the unemployment level never reaches the tipping point (which in this case is 20). The endogenous process then generates a lower unemployment level than would have been observed without social-interaction effects.[6]

The general conclusion to be drawn from these analyses is that social interactions and the endogenous processes they give rise to can influence local unemployment levels considerably, upwards as well as downwards. Whether they lead to more or less unemployment than would be expected on the basis of pure business cycle effects crucially depends upon whether or not the unemployment level ever exceeds the tipping point. The results thus suggest that social interactions, in addition to making the length of individuals' unemployment spells dependent upon the unemployment of others, are likely to generate a greater variation in unemployment levels between neighborhoods or peer groups than would have been the case in the absence of such effects.

Social interaction and youth unemployment in the Stockholm metropolitan area

Data

The dataset that we use is a population-level panel data with information on all 21- to 24-year-olds who lived in the Stockholm metropolitan area during the period from January 1992 to December 1999. We here define the Stockholm metropolitan area as consisting of the entire Stockholm County, except for the following municipalities, which are situated at the outer borders of the county: Norrtälje, Sigtuna, Upplands Bro, Södertälje, Nykvarn and Nynäshamn. The size of the remaining land area is approximately 2,616 square kilometers and the distance between the centroids of the two most distant municipalities, Vallentuna and Haninge, is approximately 48 kilometers. Given the excellent public transportation system in this area, it seems reasonable to treat this area as one within which an unemployed individual could take any job he or she was offered.

The dataset contains information on 300,619 individuals in this age group. We obtained information from various administrative registers on their demographic characteristics, including age, sex, education,

[6] With a linear benefit difference function of the form $v(t) = a + b\,u(t)$, the tipping point is equal to $-a / b$.

income and country of birth, and also information about their parents. For those who were ever unemployed we know the dates and exact lengths of all their unemployment spells measured in number of days.[7] During these years, 94,707 individuals had at least one unemployment spell during the period from January 1 the year they turned 21 to December 31 the year they turned 24.

Following in the tradition of the Swedish geographer Torsten Hägerstrand, we assume that social interactions in part reflect individuals' spatial locations: the closer two individuals are to one another, the more likely they are to be aware of and influence each other's behavior (see Hägerstrand 1965, 1967). We know where these individuals lived at the end of each calendar year, and using this information we adopt a geographically defined peer group that appears reasonable for our purposes. The Stockholm metropolitan area is divided into 619 so-called SAMS areas. These geographical areas, which have been constructed so as to contain socially homogeneous residential areas, serve as the basis for our definition of the likely peer group. The peer group consists of those 21- to 24-year-olds who resided in the same or adjacent SAMS area to that of the focal individual. The median number of 21- to 24-year-olds in these peer groups was equal to 1,082.

We restrict the analysis to 21- to 24-year-olds for two major reasons. First, by focusing on this narrowly defined age group we are likely to reduce the magnitude of unobserved heterogeneity as compared to what would have been the case had we focused on the entire labor force. Second, we focus on this age group because it is likely that their peer groups are to a large extent located in close geographic proximity.

The main reason for restricting the analysis to a single metropolitan area is that we wish to hold constant one of the most important environmental variables: the local labor market situation. Given the fairly short commuting distances within the Stockholm metropolitan area, it can, for all practical purposes, be viewed as one and the same labor market. Thus, by restricting the analysis to a single metropolitan area, we reduce the risk of mistaking spatial variations in vacancy rates and other labor market conditions for interaction-based peer-group effects.

Neighborhood variations

As the analyses reported in Figure 10.3 revealed, social interactions and the endogenous processes they give rise to can influence local

[7] We focus on "open" unemployment, which means that we do not consider those engaged in education or labor market training programs to be unemployed.

unemployment levels considerably, upwards as well as downwards. Some neighborhoods will have a higher unemployment level than they otherwise would have had, and other neighborhoods will have a lower level. This will show up as an excess variation in unemployment levels between neighborhoods as compared to what would have been observed without social interactions.

We will test for excess variation using a technique developed by Glaeser *et al.* (1996). The basic idea is fairly simple: if each individual has a probability p of being unemployed, each neighborhood will have an unemployment level equal to p with some variance around that value. Since p is a binomial variable we can calculate the expected variance across neighborhoods and compare this with the observed variance. If social interactions and the endogenous processes they give rise to are important, the observed between-neighborhood variation should be larger than the theoretically expected variation.

Following Glaeser *et al.* (2000) we can define a social interaction index as follows:

$$\mathrm{SII}_t = \frac{Var((p_{jt} - p_{.t}) \times \sqrt{n_{jt}})}{p_{.t} \times (1 - p_{.t})}$$

where $Var()$ is the variance, p_{jt} refers to the proportion unemployed in area j at time t, $p_{.t}$ is the proportion unemployed in the entire Stockholm metropolitan area at time t, and n_{jt} is the number of individuals in area j at time t. If the observed unemployment level in a particular area is equal to $p_{.t}$ plus some random noise, we should expect this index to be close to 1.0. As can be seen from the uppermost line in Figure 10.4, the observed values are considerably higher, however. The average value of the social-interaction index is equal to 2.84 (the upper limit of the 5 percent confidence interval varies between 1.0115 and 1.0120, and is shown as the almost straight line at the bottom of the graph).

The fact that the unadjusted indices are much greater than 1 should not come as a surprise to anyone since the distribution of individuals into different neighborhoods is not the outcome of random assignment. Individuals residing in different neighborhoods differ systematically from one another with respect to such factors as education and ethnicity that influence their labor market opportunities. To control for such differences, we estimated 96 OLS regressions, one for each month, where the dependent variable indicated whether an individual was unemployed or not on the 15th of each month, and the independent variables measured the individual's age, sex, education, marital status, number of children, country of birth, whether the individual was a student,

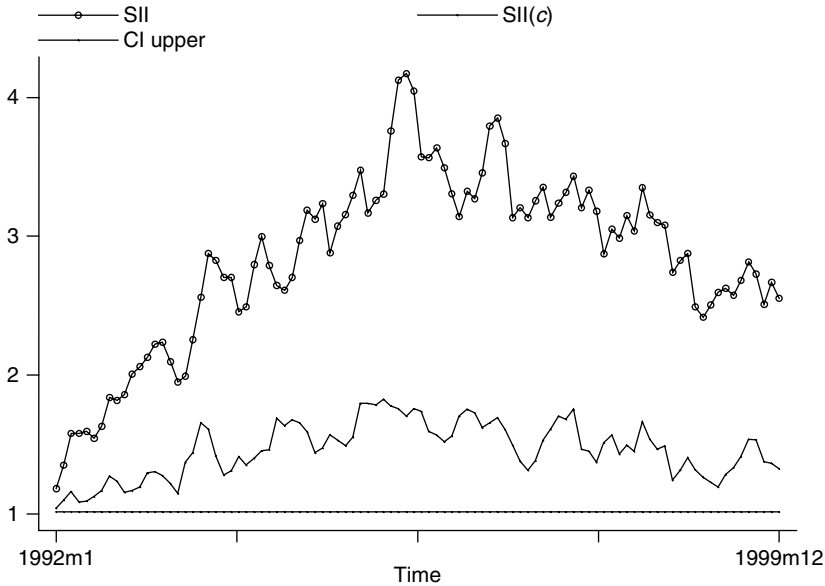


Figure 10.4 Social interaction indices before and after regression controls for individual-level differences.

and whether he/she was a recent immigrant. The *residuals* from these analyses were then used to examine the variation in unemployment levels across neighborhoods after controlling for between-neighborhood heterogeneity as measured with this set of independent variables. The social interaction index with controls for these compositional differences was then calculated as follows:

$$\mathrm{SII}(c)_t = \frac{Var((r_{jt} - r_{.t}) \times \sqrt{n_{jt}})}{Var(r_{it})}$$

where r_{jt} refers the average residual in area j at time t, $r_{.t}$ is the average residual in the entire Stockholm metropolitan area at time t, and $Var(r_{it})$ is the variance of the residuals in the entire Stockholm metropolitan area at time t.

The adjusted social interaction indices are also shown in Figure 10.4. Even after these rather extensive controls, the average value of the adjusted interaction index is 1.46. This means that, on average, the between-neighborhood variance was 1.46 times larger than one would have expected it to be given the between-neighborhood differences in age, sex, education, etc.

Qualitatively, these results are in line with our theoretical expectations. As Figure 10.3 suggested, endogenous processes are likely to produce excess variation between neighborhoods. In the Stockholm metropolitan area, the unemployment level in this age group varied from 7 percent in January 1992, to a high 17 percent in August 1993, and to a low 3 percent in December 1999. If endogenous interaction effects were at work we should expect to find a pattern very similar to that in Figure 10.4, i.e. low excess variance at the beginning and at the end of the 1990s, and substantial excess variance in the mid-1990s.

Peer group effects

In this section, we do not focus on the aggregate outcomes, but on the interactions as such, i.e. on whether or not the unemployment of peers reduces an individual's likelihood of leaving unemployment.

In the model analyzed earlier in the chapter (see pp. 208–10), we assumed that the rate at which unemployed individuals leave unemployment is proportional to the benefit difference between being unemployed and employed, and that the benefit difference is, in turn, a function of the number of unemployed in the peer group. This idea can be expressed as follows:

$$\Pi_{jt} = \omega + \gamma \times u_{jt-1} \tag{2}$$

where Π_{jt} is the conditional probability per unit of time that an individual in group j will leave unemployment at time t given that the individual was unemployed at t, and u_{jt-1} is the proportion of unemployed in group j at t-1.

The theoretical model discussed earlier (pp. 208–10) assumed the existence of groups of homogeneous individuals, but this assumption will not do when focusing on real individuals. Individuals differ in terms of their age, education, sex, ethnicity, etc., and this means that their benefit functions will differ as well. Some individuals have more remunerative alternatives to unemployment than do others, and this means that they will have higher ω-values. Similarly, some individuals may be more susceptible to social influence from peers than others, and this means that they will have more negative γ-values.

The variation in ω- and γ-values can be modeled as a function of relevant attributes of the individuals:

$$\omega = a + gZ_{it-1} \tag{3}$$

$$\gamma = c + qY_{it-1} \tag{4}$$

where

Z_{it-1} = a set of covariates describing individual i at time t-1,

Y_{it-1} = a set of covariates describing individual i at time t-1 (which may or may not overlap with Z_{it-1}).

Combining Equations (2), (3), and (4) (and allowing for an individually specific error term e_{ijt}) we get:

$$\begin{aligned}\Pi_{ijt} &= \omega_{it-1} + \gamma_{it-1}u_{jt-1} = (a + gZ_{it-1} + v_{it}) + (c + qY_{it-1} + w_{it})u_{jt-1} \\ &= a + cu_{jt-1} + gZ_{u-1} + qY_{it-1}u_{jt-1} + e_{ijt}\end{aligned} \tag{5}$$

It is the parameters of this type of equation that we will estimate. The specific model that we will estimate is a Cox proportional hazards model of the following type:

$$h_{it} = h_{0it}e^{cu_{jt-1} + gZ_{it-1} + qY_{it-1}u_{jt-1} + mN_{jt} + sT_t + rD_j} \tag{6}$$

where h_{it} is the hazard of individual i leaving unemployment at time t, u_{jt-1} is the unemployment level in the peer group in neighborhood j at time t-1, Z_{it-1} and Y_{it-1} are two sets of covariates describing individual i at time t-1, N_{jt} is a set of variables describing neighborhood j at time t, T_t is a set of monthly and annual dummy variables, and D_j is a set of dummy variables identifying which neighborhood the individual resided in.

The key variable is the peer-group unemployment variable, u. It is equal to the proportion unemployed within an individual's peer group on the 15th of each month (not including the focal individual). The main purpose of the analysis is to examine whether this variable is systematically related to the unemployed individual's hazard of leaving unemployment during the subsequent month when controlling for other relevant factors. The estimates are found in Table 10.1.

The first model in Table 10.1 relates the hazard of leaving unemployment to the unemployment level in the peer group. The hazard ratio is less than 1.0, which means that the more unemployed there are in the peer group, the lower is the hazard of leaving unemployment. The value of 0.034 suggests a substantial peer group "effect." Making an out-of-sample prediction, it suggests that if everyone else in the peer group were unemployed, the individual's chances of leaving unemployment would only be about 3 percent of what they would have been had no one been unemployed. Obviously, much of this peer group "effect" is due to labor market fluctuations and individual

Table 10.1. *Cox regression, hazard ratios of leaving unemployment (z statistics in parentheses)*

	Model 1	Model 2	Model 3	Model 4	Model 5
Prop. unemployed in peer group	0.034*** (-66.16)	0.123*** (-29.03)	0.083*** (-20.41)	0.288*** (-8.868)	0.345*** (-5.927)
Vacant positions in Stockholm County		1.083*** (15.57)	1.082*** (15.48)	1.085*** (15.94)	1.091*** (13.66)
Women		1.144*** (28.98)	1.146*** (29.16)	1.147*** (29.21)	1.166*** (25.86)
Age		0.980*** (-9.232)	0.981*** (-8.942)	0.979*** (-9.456)	0.981*** (-6.74)
High school education		1.105*** (18.79)	1.111*** (19.67)	1.117*** (20.49)	1.118*** (15.90)
College education		1.167*** (19.15)	1.175*** (19.97)	1.181*** (20.43)	1.165*** (15.65)
From EU outside Sweden		1.021 (1.411)	1.021 (1.428)	1.019 (1.284)	0.959* (-1.906)
From Eastern Europe or former Soviet Union		0.946*** (-3.290)	0.943*** (-3.476)	0.945*** (-3.332)	0.968 (-1.217)
From Middle East or Africa		0.905*** (-8.031)	0.911*** (-7.523)	0.921*** (-6.520)	0.958** (-2.244)
From America		1.087*** (5.957)	1.088*** (5.993)	1.091*** (6.134)	1.086*** (4.035)
From the rest of the world		1.014 (0.803)	1.016 (0.903)	1.027 (1.518)	1.086*** (3.210)
Less than three years in Sweden		0.681*** (-22.40)	0.677*** (-22.69)	0.680*** (-22.36)	1.017 (0.313)
Three to five years in Sweden		0.959*** (-3.005)	0.958*** (-3.055)	0.957*** (-3.144)	1.071* (1.839)
Married		0.977** (-2.064)	0.974** (-2.293)	0.971** (-2.519)	0.934*** (-3.303)
No. of children		0.945*** (-6.840)	0.941*** (-7.324)	0.943*** (-7.024)	0.967** (-2.066)
Amount social welfare		0.984*** (-8.612)	0.984*** (-8.266)	0.984*** (-8.313)	0.971*** (-8.201)
Amount of sick allowance		0.956*** (-17.04)	0.955*** (-17.33)	0.954*** (-17.63)	0.950*** (-13.22)
Days unemployed before current period		0.994*** (-15.74)	0.993*** (-16.18)	0.994*** (-14.50)	0.993*** (-13.18)
Info. missing on both parents		0.951*** (-3.461)	0.963*** (-2.648)	0.960*** (-2.801)	0.995 (-0.226)
Parent(s) college educated		1.032*** (5.926)	1.041*** (7.483)	1.042*** (7.509)	1.037*** (5.551)

Table 10.1. (*cont.*)

	Model 1	Model 2	Model 3	Model 4	Model 5
Both parents from outside EU		0.925*** (-7.263)	0.929*** (-6.802)	0.931*** (-6.577)	0.886*** (-8.173)
One parent from outside EU		0.968*** (-3.391)	0.972*** (-2.964)	0.971*** (-3.017)	0.932*** (-5.649)
Parent's income (highest earner)		1.001*** (5.162)	1.001*** (6.140)	1.001*** (6.234)	1.001*** (4.559)
Parent(s) unemployed		1.001 (0.117)	1.000 (0.0663)	0.999 (-0.143)	0.992 (-1.044)
Parent(s) social assistance		0.984*** (-2.712)	0.982*** (-2.920)	0.981*** (-3.174)	0.979** (-2.519)
Prop. college educ. in peer group			0.885* (-1.900)	1.011 (0.0910)	0.732** (-2.099)
Prop. born outside EU in peer group			0.831* (-1.753)	0.521*** (-3.218)	0.477*** (-2.688)
Mean time unemployed among adult neighbors			0.925*** (-2.835)	0.598*** (-9.230)	0.571*** (-7.538)
Prop. college educated among adult neighbors			0.947 (-0.934)	0.338*** (-2.681)	0.834 (-0.347)
Prop. born outside EU among adult neighbors			1.010 (0.0894)	2.299*** (2.905)	3.554*** (3.215)
Mean income among adult neighbors			0.953*** (-5.855)	1.014 (0.596)	1.006 (0.188)
Grade point average					1.063*** (12.38)
Dummies for years and months	No	Yes	Yes	Yes	Yes
Dummies for neighbourhoods (= fixed effects)	No	No	No	Yes	Yes
Number of individuals	94,707	94,707	94,707	94,707	61,417
Log likelihood	-2187330	-2171766	-2171657	-2170411	-1325150

*** p<0.01, ** p<0.05, * p<0.1 (two-sided test).

heterogeneity across neighborhoods, and we will gradually introduce controls for this.

In the second model, we include monthly and annual dummy variables to control for seasonal variations and time trends,[8] and we control for the number of job vacancies that were available at the beginning of each month in the Stockholm County (according to statistics from the Swedish labor market authorities). We also introduce various variables to control for relevant differences between the individuals residing in different neighborhoods: sex, age, education (highest degree), country of birth, number of years residing in Sweden, marital status, number of children, amount of social welfare and sick allowance received during the previous calendar year, and previous unemployment experiences measured as the total number of unemployment days before the current unemployment period started. The variable measuring previous unemployment experiences has been included in order to control for unobserved and otherwise uncontrolled heterogeneity likely to influence the hazard of an individual leaving the current unemployment spell. The length of the current unemployment spell is controlled for in the baseline hazard.

We also introduce variables describing basic characteristics of the parents of the individuals. We have information about parents for about 90 percent of the individuals; the missing information is due to cases where the father was unknown, and where parents lived abroad or no longer were alive. We include variables describing the parents' education, country of birth, income, and their unemployment experiences and receipt of social assistance during the previous year.

The most important result in Model 2 is that the unemployment level in the peer group has a substantial effect on the hazard even after we control for all of these individual and parental attributes, and for temporal variations in labor market conditions. The hazard ratio of 0.123 suggests that if everyone else in the peer group were unemployed, the individual's risk of leaving unemployment would be only about 12 percent of what it would have been had no one been unemployed. But, once again, this is an out-of-sample prediction and should therefore be viewed with some caution. (The effects of the control variables are discussed below.)

This peer-group effect could, however, be spurious and due to other peer group and neighborhood characteristics that are correlated with the unemployment level among the peers. In Model 3 we therefore introduce covariates that describe other properties of the peers and

[8] To save space, we have not included these estimates here, but they are available from us upon request.

the neighborhoods. We control for the proportion of college education among the peers and the proportion of the peers who were born outside of the EU. We also control for the same factors for neighbors of prime working age, 30–60 year olds, and for the average number of unemployment days and average income during the year within this older age group. When including these neighborhood controls, the effect of the peer group unemployment level becomes stronger, not weaker, and the hazard ratio decreases from 0.123 to 0.083. This is a strong indication that the unemployment level in the peer group has an influence in its own right on the transition rate out of unemployment, and that it is not just a spurious effect.

The fourth model is equivalent to Model 3, but it is a fixed-effect specification including 618 dummy variables, one for each neighborhood (except one). The reason for including these dummy variables is to control for all time-invariant unobservable characteristics of the neighborhoods. This way of controlling for between-neighborhood differences may mean that we introduce excessive controls and therefore underestimate the true effect of the unemployment level among the peers, since these dummy variables may absorb some social-interaction effects that are stable and long-lasting. But even with these extensive controls the hazard ratio associated with the peer group unemployment variable is as low as 0.288, suggesting a most substantial social-interaction effect.

The fifth model is identical to the fourth model, except for the fact that we only use data on individuals with a high school or vocational school diploma in order to be able to include their grade point average as an additional control variable.[9] Controlling for grade point average and restricting the analysis to this sub-population somewhat reduced the effect of the unemployment level among peers, but the effect remains considerable; the hazard ratio is equal to 0.345.

The effects of the covariates are also interesting, but they are not our primary concern in this chapter. We will therefore briefly mention only a few of them. The results referring to the properties of the individuals and the properties of the parents are mostly what one would expect considering previous research.

The effects of the variables describing the peer groups and the larger neighborhood community differ considerably between the models with and without neighborhood dummies. Model 3 is the most informative model if one is interested in how variations *between* neighborhoods

[9] During these years, grades in Swedish high schools varied from a low of 1 to a high of 5.

influence the effect on the transition rate out of youth unemployment, and we therefore limit our discussions to Model 3.

As expected, a higher unemployment level among 30–60 year olds in the neighborhood decreases the chances of leaving unemployment. The most interesting results refer to the effects of the income of those of prime working age within the neighborhood. According to Model 3, the transition rate out of unemployment is inversely related to the average income of the 30–60 year olds within the neighborhood. This unexpected result may reflect socially conditioned expectations or aspiration levels which make young persons in rich neighborhoods choosier about which jobs they accept.

The model in Table 10.2 examines whether individuals with certain characteristics appear to be more susceptible to influence than others. It does this by examining statistical interaction effects between the unemployment level in the peer group and various demographic variables. The model also includes all the other variables from Model 4.

To make the statistical interaction effects in Table 10.2 easier to interpret, they are graphed in Figure 10.5. The solid lines in all these graphs represent the reference category, which consists of 21-year-old men with education lower than high school, who are not recent immigrants. The hazard ratio is equal to 1 when there is no unemployment among the peers. Each graph shows how the hazard ratio varies with a certain individual characteristic and with the unemployment level in the peer group.[10]

The first graph shows that women leave unemployment faster and are less influenced by the unemployment level among their peers than are men. The second graph shows that the slightly older have a harder time leaving unemployment and are less influenced than are the slightly younger ones. The third graph shows that the less educated are more influenced by peers than the more educated are, and that it tends to take a longer time for them to leave unemployment. Figure 10.5 also shows that persons who have lived less than three years in Sweden have a hard time leaving unemployment and that those who immigrated to Sweden three to five years ago are less influenced by the unemployment level among their peers than are those who have lived in Sweden for more than five years. It is especially worth noting that when the unemployment level in their peer group is lower than about 17 percent they have lower chances of leaving unemployment than those who have lived a longer time

[10] The graph for the reference group is calculated by the equation: $hazard\ ratio = c^u$, where u is the proportion unemployed in the peer group, and c is the hazard ratio for

Table 10.2 *Cox regression, hazard ratios of leaving unemployment (z statistics in parentheses), with statistical interaction effects*

Prop. unemployed in peer group	0.147***
	(-11.59)
Women	1.121***
	(10.07)
Age	0.965***
	(-6.864)
High school education	1.065***
	(4.954)
College education	1.039*
	(1.944)
Less than 3 years in Sweden	0.568***
	(-15.95)
3–5 years in Sweden	0.854***
	(-5.258)
Prop. unemployed in peer group * Women	1.261**
	(2.370)
Prop. unemployed in peer group * Age	1.155***
	(3.227)
Prop. unemployed in peer group * High school educ.	1.548***
	(3.972)
Prop. unemployed in peer group * College educ.	3.439***
	(7.387)
Prop. unemployed in peer group * <3 years in Sweden	4.605***
	(5.930)
Prop. unemployed in peer group * 3–5 years in Sweden	2.564***
	(4.424)
All other variables from Model 4	Yes
Log likelihood	-2170344

*** p<0.01, ** p<0.05, * p<0.1.

in Sweden, but the opposite is true when the unemployment level is higher than about 17 percent.

With these data we cannot determine why we observe these differences, but it seems likely that the age-, sex- and immigration-based (statistical) interaction effects are related to how embedded the individuals are in their peer groups. The younger cohorts and their peers are likely to have lived together in the same neighborhoods for a longer time than the older cohorts because many of them have not yet left their parental homes. Similarly, those who recently arrived to Sweden have not yet built up extensive neighborhood-based networks. Finally, the education-based interaction effect may indicate that neighborhood-based networks are more important for finding jobs for those with less education.

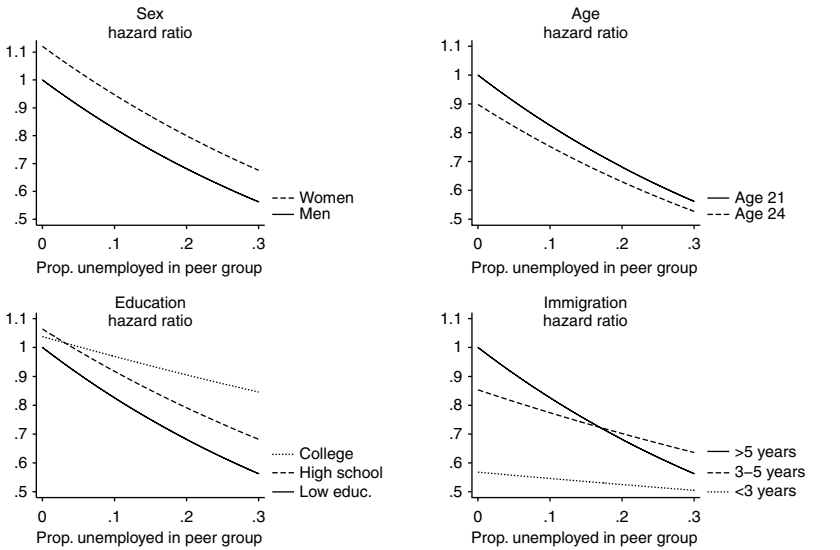


Figure 10.5 Hazard ratios for leaving unemployment with statistical interaction effects between individual attributes and the unemployment level in the peer group.

Figure 10.5 also gives some indication of the "relative importance" of the peer-group effect. In most of the graphs, there is a larger difference in the hazard ratio between peer groups with different unemployment levels than there is between individuals with different individual-level characteristics, which would suggest that the peer-group effects are mostly larger. Another useful comparison point is a typical variation in the unemployment level as measured by the standard deviation. This standard deviation is equal to about 0.047. The results reported in Model 4, which includes dummies for neighborhoods, suggest that an increase in the unemployment level with 0.047 units will tend to reduce the hazard by approximately 6 percent, and the results in Model 3, which does not include any neighborhood dummies, suggest that the hazard would be reduced by about 11 percent. Typical standard-deviation changes in most of the other covariates lead to smaller changes in the hazards.

In Figure 10.6 we compare the effect of the unemployment level in the neighborhood before and after introducing these various controls.

the proportion unemployed in the peer group. The graphs for the other groups are calculated by the equations: *hazard ratio* = $c^u g q^u$, where g is the hazard ratio for the relevant individual dummy variable, and q is the hazard ratio for the relevant statistical interaction variable.

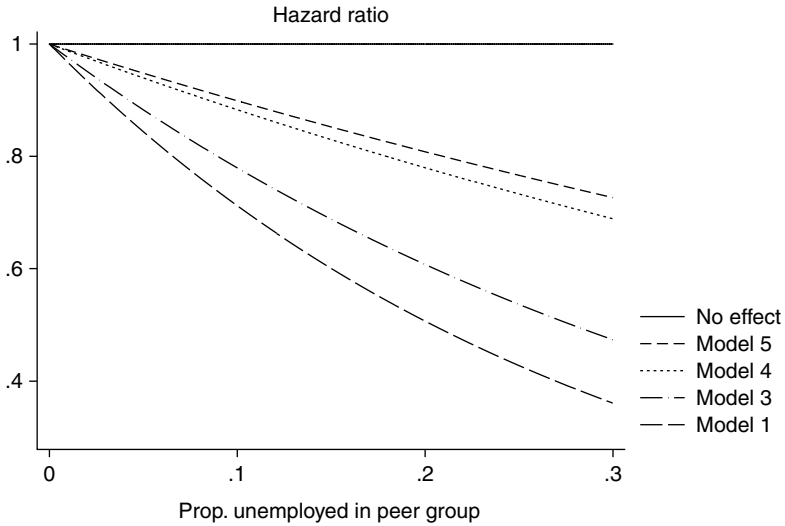


Figure 10.6 The effect of peer group unemployment on the hazard ratios for leaving unemployment, before and after controls for confounding variables.

The graphs are based on the results in Models 1, 3, 4 and 5. As can be seen from the figure, the reduction in the effect is rather substantial, but the remaining effect is even more substantial. Introducing additional control variables is likely to reduce the effect even further, but it seems highly unlikely that an effect of this magnitude exclusively or even largely could be due to omitted variables (at least we cannot imagine what variables that might be).

All in all, these results strongly suggest that the unemployment level in the peer group considerably influences the chances of an unemployed individual leaving unemployment. Although some of these peer group effects are likely to be due to remaining and uncontrolled for differences between individuals residing in neighborhoods with different unemployment levels, it seems highly unlikely that such factors could wipe out these rather substantial peer-group effects.

Conclusion

In this chapter, we have analyzed unemployment among 21- to 24-year-olds in the Stockholm metropolitan area, emphasizing the potential importance of social interactions and endogenous processes. In this

age group, about every third person experienced at least one period of unemployment during the 1990s. When being unemployed becomes such a common phenomenon, it is likely that whatever stigma used to be attached to being unemployed largely disappears, that unemployed individuals reduce their job search because they believe that they will not find any jobs, and that the networks through which they used to find jobs no longer function as well as they once did. In such situations endogenous self-reinforcing processes can exert considerable influence on local unemployment levels.

The empirical analyses revealed an excess variation in unemployment levels across neighborhoods that was patterned as one would expect it to be if social interactions were at work. We also found that individual transition rates out of unemployment were closely related to the unemployment levels in their peer groups, a result to be expected if social interactions were at work. The results therefore suggest that social interactions and endogenous processes most likely were of considerable importance for the temporal and spatial variation in youth unemployment observed in the Stockholm metropolitan area during the 1990s.

As emphasized by Crane (1981), endogenous processes such as those considered here have important policy implications. If such processes are at work, it may be advisable to concentrate public resources disproportionately in neighborhoods with high unemployment levels. If a policy intervention manages to reduce the unemployment level in a neighborhood below its tipping level, the endogenous process is likely to lead to a much greater reduction in unemployment than if the same resources were distributed more evenly across neighborhoods. The existence of endogenous processes also suggests that the effects of policy interventions may be highly non-linear. Large interventions that do not bring the unemployment level below the tipping point may be counteracted and neutralized by the endogenous process focused upon here, while small interventions may have huge effects if they bring the unemployment level below the tipping point.

In this chapter we have focused exclusively on youth unemployment, but the types of process that we have analyzed operate in numerous other areas of social life as well. Social interactions and the endogenous processes they give rise to are at the core of several key works of modern sociology such as Merton's ([1948] 1968) analysis of self-fulfilling processes, Coleman *et al.*'s (1957) analysis of network diffusion, and Granovetter's (1978) work on threshold-based behavior. We believe that the generality of sociological theory is to be found at this level of semi-general social mechanisms that operate according to the same principles in a range of social settings.

Endogenous processes such as these are, for instance, important for explaining such widely divergent phenomena as the increase in divorce rates during the second half of the twentieth century (Åberg 2009) and the growth of social movements in nineteenth-century Sweden (Hedström 1994). For sociological theory to be of real explanatory use, it must provide a toolbox of semi-general mechanisms such as those focused upon here from which precise middle-range theories can be constructed.

REFERENCES

Åberg, Y. 2000. "Individual social action and macro-level dynamics: a formal theoretical model," *Acta Sociologica* 43(3): 193–205.

2009. "Contagious divorces," in P. Hedström and P. Bearman (eds.), *The Oxford Handbook of Analytical Sociology*. Oxford University Press, ch. 15.

Clark, A.E. 2001. "Unemployment as a social norm: psychological evidence from panel data," *The Journal of Labor Economics* 21(2): 323–51.

Coleman, J.S., E. Katz and H. Menzel. 1957. "The diffusion of an innovation among physicians," *Sociometry* XX: 253–70.

Crane, J. 1981. "The epidemic theory of ghettos and neighborhood effects on dropping out and teenage childbearing," *American Journal of Sociology* 96: 1226–59.

Elster, J. 1989. *The Cement of Society: A Study of Social Order*. Cambridge University Press.

Glaeser, E.L., B. Sacerdote and J.A. Scheinkman. 1996. "Crime and social interactions," *The Quarterly Journal of Economics* May: 507–48.

Glaeser, E.L., D.M. Cutler and K. Norberg. 2000. "Explaining the rise in teenage suicide," NBER Working Paper No. w7713.

Granovetter, M. 1978. "Threshold models of collective behavior," *American Journal of Sociology* 83: 1420–43.

1995. *Getting a Job: a Study of Contacts and Careers*. University of Chicago Press.

Hägerstrand, T. 1965. "A monte carlo approach to diffusion," *Archives Européennes de Sociologie* VI: 43–67.

1967. *Innovation Diffusion as a Spatial Process*. University of Chicago Press.

Hedström, P. 1994. "Contagious collectivities: on the spatial diffusion of Swedish trade-unions, 1890–1940," *American Journal of Sociology* 99: 1157–79.

2005. *Dissecting the Social: On the Principles of Analytical Sociology*. Cambridge University Press.

Hedström, P. and P. Bearman (eds.) 2009. *The Oxford Handbook of Analytical Sociology*. Oxford University Press.

Hedström, P., K.-Y. Liu and M.K. Nordvik. 2008. "Interaction domains and suicide: a population-based panel study of suicides in Stockholm, 1991–1999," *Social Forces* 87(2): 713–40.

Manski, C.F. 2000. "Economic analysis of social interactions," *Journal of Economic Perspectives* 14(3): 115–36.

Merton, R.K. [1948] 1968. "The self-fulfilling prophecy," in *Social Theory and Social Structure*. New York: The Free Press, 475–90.

Sampson, R.J., J.D. Morenoff and T. Gannon-Rowley. 2002. "Assessing 'neighborhood effects': social processes and new directions in research," *Annual Review of Sociology* 28.

Schelling, T.C. 1978. *Micromotives and Macrobehavior*. New York: W.W. Norton.

Schweitzer, A.O. and R.E. Smith. 1974. "The persistence of the discouraged worker effect," *Industrial and Labor Relations Review* 27(2): 249–60.

Weber, M. [1921–2] 1978. *Economy and Society*. Berkeley: University of California Press.

Wilson, W.J. 1987. *The Truly Disadvantaged: the Inner City, the Underclass, and Public Policy*. University of Chicago Press.

Zawadski, B. and P. Lazarsfeld. 1935. "The psychological consequences of unemployment," *Journal of Social Psychology* VI: 224–51.

11 Neighborhood effects, causal mechanisms and the social structure of the city

Robert J. Sampson

The idea of "neighborhood effects" has emerged as a sharp point of contention in the social sciences. Although most scholars would probably agree that important aspects of life are disproportionately concentrated by place and that spatial inequality looms large in cities around the world, disagreement reigns over the meaning of such facts (Sampson 2008). Indeed, disputes have erupted across multiple disciplines over the proper level of analysis for assessing neighborhood effects, the role of selection bias, the social mechanisms at work, proper methods of measurement, and ultimately the nature of causal inference in a social world.

The stakes of this debate are high given the wide range of behaviors potentially subject to neighborhood effects. A large literature demonstrates that concentrated inequality covers a diverse array of phenomena, including but not limited to crime, economic self-sufficiency (e.g. unemployment, welfare use), asthma, infant mortality, depression, violence, drug use, low birth weight, teenage pregnancy, cognitive ability, school dropout, child maltreatment and even Internet use, to name but a few.[1] Contrary to the fashionable claim of the "death of distance," concentrated inequality is a durable presence despite globalization (Sampson 2011).

The study of neighborhood effects therefore makes a good case for reflecting on social mechanisms and causal processes. I do just that in an attempt to advance analytic thinking on the social mechanisms that constitute neighborhood effects. There are many entry points to the debate and one cannot make scientific headway on all of them in one chapter.[2] For pragmatic reasons I focus on issues that I believe

[1] I set aside a review of the by-now voluminous evidence supporting this general claim. Jencks and Mayer (1990) provided an extensive evaluation of existing research on the effects of growing up in neighborhood poverty. More recent and general syntheses of neighborhood effects are in Leventhal and Brooks-Gunn (2000), Sampson *et al.* (2002) and Sampson (2011).

[2] This chapter draws on a larger effort that attempts to make such headway (Sampson 2010). See also Wikström and Sampson (2003, 2006).

bear most directly on the concerns of "analytic sociology" (Hedström 2005; Hedström and Bearman 2009). After a conceptual definition of neighborhood units of analysis, I discuss emergent processes (vs. social composition or the aggregation of individual attributes), a strategy for neighborhood-level measurement ("ecometrics"), and the implications of social mechanisms that operate at the "extra-local" level – beyond the neighborhood as it were. I illustrate these points by reference to research that conceptualizes and assesses a theory of collective efficacy as an explanation of variations in crime rates.

I then turn to so-called "selection bias" and how resulting connections among neighborhood units arising from mobility decisions (a form of individual action) constitute the social structure of the contemporary city. The residential selection of individuals is of particular concern to the burgeoning literature on neighborhood effects. Because individuals make choices based on preferences and beliefs (within their means, of course), they are said to allocate themselves non-randomly. As a result it is often thought that estimates of neighborhood effects are confounded and thereby biased. The main response is to view selection bias as a statistical problem to be controlled away rather than something of substantive interest in itself.

By contrast, I consider selection as a mechanism that is:

1. itself influenced or caused by neighborhood factors; and
2. generates potentially important implications for inequality in neighborhood-level attainment and broader social-level processes.

I specifically focus on selection in the form of neighborhood sorting, where I treat the neighborhood attainment of an individual as problematic in its own right and requiring explanation. This in turn motivates a focus on the social consequences of residential selection. Here the question becomes how individual decisions combine to create spatial flows that define the ecological structure of inequality, an example of what Coleman (1990: 10) more broadly argued is a major under-analyzed phenomenon – micro-to-macro relations. Building on recent work I translate these concerns to an analysis of structural flows of exchange between neighborhoods of different racial and economic status that are fundamental to the social reproduction of racial inequality.

Definitions and background facts

I begin with a general definition of neighborhood as a variably interacting population of people and institutions in a common place. Neighborhoods form a mosaic of overlapping ecological units (e.g. blocks, streets) that

vary in size, boundaries and social organizational features.[3] Grounded in a larger systemic theory of urban residential differentiation (Sampson 1988), my strategy begins with neighborhoods in physical space rather than elevating social interactions or identity to the definitional criteria. As Warren argued over thirty years ago, belief in the demise of neighborhood as an important social unit is predicated on the assumption that neighborhoods are a primary group that possess the "face-to-face" intimate, affective relations which characterize all primary groups (Warren 1975: 50). But the extent of structural or cultural organization (if any, and for what) is an empirical question subject to scrutiny.

A logical implication is that sometimes neighborhoods make a community in the traditional sense of shared values and tight-knit bonds, but many times they do not. When formulated in this way, factors such as network density, attachment to place, civic participation, disorder, organizational density, identity and capacity for collective action are *variable* and analytically separable not only from potential structural antecedents (e.g. economic resources, ethnic diversity, residential stability) and possible consequences (e.g. crime, innovation), but from the definition of the units of analysis. I thus define neighborhoods geographically and leave the nature and extent of social relations problematic (Tilly 1973: 212). This conceptualization opens the door for empirical research to proceed without tautology and a menu of ecological units of analysis from which to choose depending on the theoretical constructs or social phenomenon under study.

My approach contrasts with the idea that there should be one "correct" or invariant operational definition of neighborhood and likewise the associated belief that because neighborhood boundaries are differentially perceived (even among individuals living in the same physical setting), neighborhood is not a valid theoretical category. How this argument goes wrong can be appreciated by considering the analogy to invariant conceptions of family, church or nation. How family is defined is variable and the subject of fierce debate, while churches take on a strikingly heterogeneous form (from a handful of people in a living

[3] In practice, most empirical studies rely on geographic boundaries defined by governmental agencies (e.g., the US Census Bureau; local school districts, city police districts). Although administratively defined units such as census tracts in the United States or political wards in the UK are reasonably consistent with the notion of overlapping and nested ecological structures, researchers have become increasingly interested in strategies to define neighborhoods that respect the logic of street patterns and the social networks of neighborly interactions (Grannis 1998) and the activity patterns of individuals (Wikström *et al.* 2010). Sampson *et al.* (2002) discuss the pros and cons of different choices, and the fortunate outcome that most results are robust across definitional units, suggesting a general form to neighborhood effects.

room to 20,000-strong mega-congregations). Or consider the varying and time-dependent definitions applied to nation states. Is nation (or church or family) then simply a social construct void of causal power? Must people within a nation agree on its definition or its borders? Few social scientists would infer that societies lack causal power because their boundaries are socially constructed, permeable and variable. As Brubaker (1996) argues, *nationalism* is a legitimate object of causal scientific inquiry that is rooted in everyday practices and institutionalized forms, even if "nation" is a socially contingent concept.

Applied to the present case, we can fruitfully conceptualize categories of practice associated with a place that take on variable institutionalized forms (e.g. real-estate steering, school boundaries, community organizations). "Neighborhood-ness" (e.g. as reflected in spatially bounded social interactions, place identity, or social controls) is a contingent or variable event; neighborhoodness or community – defined by the social features associated with location – is not the same thing as a physical neighborhood. It is the intersection of practices and social meanings with spatial context that is at the root of neighborhood effects.

Beyond composition

Although concern with neighborhood social mechanisms and processes goes back to the early Chicago School, only recently have we witnessed a concerted attempt to theorize and empirically measure the social-interactional and institutional dimensions that explain how neighborhood effects are transmitted. Elsewhere my colleagues and I reviewed what we called the "process turn"' in neighborhood effects research (Sampson *et al*. 2002). Unlike the more static features of socio-demographic composition (e.g. race, class background), social processes and mechanisms provide accounts of *how* neighborhoods bring about a change in a phenomenon of interest. I conceptualize a social mechanism as a plausible contextual process that accounts for a given phenomenon, in the ideal case linking putative causes and effects (Sørensen 1998: 240; Wikström and Sampson 2003). Mechanism as explanation is thus largely a theoretical claim – mechanisms can rarely be observed or manipulated causally in an experiment. Rather, social mechanisms make up the hypothesized link in the pathway of explanation from a manipulable cause (Woodward 2003) to an outcome. The goal is to develop indicators of the sets of practices, meanings and actions that reflect hypothesized mechanisms, for example reciprocated exchange among neighbors, intergenerational closure, and social control. My research has attempted to examine:

1. the sources of these and other emergent social processes that vary across neighborhoods;
2. accounts of how social mechanisms relate to rates of social behavior, such as crime rates; and
3. contextual effects of neighborhoods on individuals.

A focus on neighborhood-level effects should not be read to imply a neglect of cultural and symbolic processes or a search for universal covering laws. My approach allows me to simultaneously probe what may be a powerful role for neighborhoods in the contemporary city – social modes of perceptual organization (Sampson 2009; Wikström 2010). Neighborhoods are usually conceptualized in terms of "structural" variables such as poverty that could be implanted anywhere. When mediating factors are studied directly, for example network ties, they too are usually considered without reference to the perceptions and interpretations that give them meaning. But as scholars have long argued, places have symbolic and not just use value (Firey 1947; Hunter 1974; Suttles 1972). Consideration of the cultural processes that make places meaningful and perceptually distinct is therefore a central task of the process turn in neighborhood effects research (Sampson 2009; Sampson *et al.* 2002).

"Ecometrics"

Unlike individual-level measurements which are backed by decades of psychometric research into their statistical properties, the methodology needed to evaluate neighborhood-based mechanisms is not widespread and deserves equal attention. Raudenbush and Sampson (1999) thus proposed moving toward a science of ecological assessment, which they call "ecometrics," by developing systematic procedures for directly measuring neighborhood mechanisms, and by integrating and adapting tools from psychometrics to improve the quality of neighborhood-level measures. Setting aside statistical details, the important theoretical point is that neighborhood processes can and should be treated as ecological or collective phenomena rather than as stand-ins for individual-level traits. I believe this distinction is crucial for the advancement of research.

Using community-level surveys and systematic social observations, Raudenbush and Sampson (1999) demonstrate reliable and valid measures of neighborhood structural, cultural and institutional processes. Oversimplifying, the evidence supports at least four classes of neighborhood social processes that, while interrelated, appear to have

independent ecometric validity for community-level theory, especially for the study of crime (Sampson *et al.* 2002).

Social networks/interaction. One of the driving forces behind much of the research on neighborhood mechanisms has been the concept of social capital, which is generally conceptualized as a resource that is realized through social relationships (Coleman 1988). The studies we reviewed included a number of measures that tap dimensions of interpersonal relations, such as density of social ties between neighbors, exchange, the frequency of social interaction among neighbors, and patterns of "neighboring." Social capital is generally thought to provide a resource through mechanisms of information exchange and interlocking ties, such as intergenerational closure between adults and children, which in turn enhance social control.

Norms and collective efficacy. Although social ties are important, the willingness of residents to intervene in neighborhood social life depends, in addition, on conditions of mutual trust and *shared expectations* among residents – a form of what Searle (1995) refers to as "we-intention." One is unlikely to intervene in a neighborhood context where the rules are unclear and people mistrust or fear one another. It is the linkage of mutual trust and the shared willingness to intervene that captures the neighborhood context of what Sampson *et al.* (1997) term *collective efficacy*. Some density of social networks is essential, to be sure, especially networks rooted in social trust. But the key theoretical point is that networks have to be activated to be ultimately meaningful. Collective efficacy theory therefore helps to elevate the "agentic" aspect of social life (Bandura 1997) over a perspective centered mainly on the accumulation of stocks of social resources as found in ties and memberships (i.e. social capital). This conceptual orientation is consistent with the redefinition by Portes and Sensenbrenner (1993) of social capital in terms of "expectations for action within a collectivity." Distinguishing between the resource potential represented by personal ties, on the one hand, and shared expectations for action represented by collective efficacy, on the other, helps clarify the dense networks paradox: *social networks foster the conditions under which collective efficacy may flourish, but they are not sufficient for the exercise of control*. This conceptualization recognizes the transformed landscape of contemporary urban life, holding that while community efficacy may depend on working trust and social interaction, it does not require that my neighbor be my friend. Sampson and colleagues (1997) constructed a measure of collective efficacy by combining informant ratings of the capacity for informal social control with social cohesion.

Another normative feature of neighborhoods, or what might be thought of as an orienting cultural climate, is *moral cynicism*. Even after

we adjust for individual characteristics such as income, race and other traditional factors, neighborhoods vary systematically in the moral cynicism of residents toward law and mutual helping behavior. In neighborhoods that have experienced concentrated disadvantage over long periods of time, a corrosive atmosphere of alienation and collectively perceived "dog-eat-dog" culture emerges, a context where law is not perceived as existentially relevant. Neighborhood "poverty traps" predict the moral cynicism of residents up to twenty-five years later, suggesting institutional processes that reinforce despair. In turn, neighborhood cynicism predicts violence (Sampson *et al*. 2005).

Organizational infrastructure. The quality, quantity and diversity of institutions constitute an important but often neglected component of neighborhood studies, especially non-profit and civic community-based organizations that provide public goods. Examples of such organizations include libraries, schools and other learning-based centers, child care centers, organized social and recreational activities, and family support centers. A key mechanism is that organizations link individuals together in unintended ways that enhance collective oriented tasks (Small 2009). A community's organizational infrastructure is correlated with but not the same thing as the density of civic participation in local organizations. In practice, however, most empirical measures have focused on the presence of neighborhood institutions based on survey reports and archival records. A few studies have used surveys and ethnography to tap levels of actual participation in neighborhood organizations and the nature of the networks that are formed. Organizational participation and institutional density have both traditionally been seen as key foundations for anti-crime and collective action efforts in neighborhood settings.

Just because an organization is located in a particular community does not mean that its interests mesh with that community, however. A study of the religious ecology of churches in Boston shows that the density of organizations is hardly sufficient, especially when the constituents of the organization (in this case, parishioners) come from *outside* the community (McRoberts 2003). In thinking about institutions theoretically, we thus need to be careful not to conflate organizational density with coordinated action for local interests. Recent research has begun to examine the link between organizational density, civic membership in organizations, and the actual mobilization of action for collective pursuits (Sampson 2011).

Activity patterns/routines. A fourth and often overlooked factor in discussions of neighborhood effects is the ecological distribution of daily routine activities. Wikström and Sampson (2003) refer to ecologically

structured routines as *behavior settings*. The location of schools, the mix of residential with commercial land use (such as strip malls, bars), public transportation nodes, and large flows of night-time visitors to the city center, for example, are relevant to organizing how and when people come into contact with others and non-resident activity. Like studies of institutions, however, direct measures of social activity patterns and behavior settings are relatively rare. More common in studies of routine activities are measures of land use, such as the presence of schools, stores and shopping malls, motels and hotels, vacant lots, bars, restaurants, gas stations, industrial units and multi-family residential units (Sampson *et al.* 2002). In short, the spatial organization of routine activities and everyday behavior settings permits a variety of social interactions (positive as well), and is therefore relevant to explaining social behavior (Wikström 2010; Wikström *et al.* 2010).

"Extra local" processes and the larger social order

Prior research on neighborhood effects has fixated on the idea of "contained" or internal characteristics, almost as if neighborhoods are islands unto themselves. This gap in our knowledge is surprising given that a traditional workhorse of urban ecology is spatial interdependence. The recent increase in the power of Geographic Information Systems (GIS) integrated with spatial analytic techniques has led to a new generation of research focusing on the interdependence of social processes through spatial networks, especially mechanisms such as diffusion and exposure (Morenoff *et al.* 2001). For example, after accounting for measured compositional characteristics internal to a neighborhood, the collective efficacy and violence in a given neighborhood are significantly and positively linked to the collective efficacy and violence rates of surrounding neighborhoods, respectively (Morenoff *et al.* 2001; Sampson *et al.* 1999). This finding suggests a diffusion or exposure-like mechanism, whereby violence and collective efficacy are conditioned by the characteristics of spatially proximate neighborhoods, which in turn are conditioned by adjoining neighborhoods in a spatially linked process that ultimately characterizes the entire metropolitan system. The mechanisms of racial segregation are a major force in such spatial inequality, explaining why it is that despite similar income profiles, black middle-class neighborhoods are at greater risk of violence than white middle-class neighborhoods (Sampson *et al.* 1999).

But even spatial thinking, although welcome, has been limited to a focus on how *internal* neighborhood characteristics are associated with the internal characteristics of a "neighborhood's neighbors" – spatial

proximity or geographic distance has been the defining metric. As a result, the ways in which cross-neighborhood networks are tied into the social structural web of the city are not well understood and virtually never studied empirically. This is again surprising, for the classic work of Park and Burgess in *The City* (1925) envisioned research on the city's ecological structure whereby neighborhoods were pieces of the larger social whole. The bottom-line message is that studies of place should not proceed by considering *indigenous* qualities only.

It is not just city-level processes that are at stake – national and global forces can influence place stratification. As Wilson (1987) argued in *The Truly Disadvantaged*, deindustrialization and the shift to a service economy was disproportionately felt in the inner city, contributing to an increasing concentration of poverty in the 1970s and 1980s. The problem is that serious empirical work at this more "bird's-eye" level is difficult. How do we go about documenting the extra-local layers of this kind of "macro-level" neighborhood effect? In fact, it might well be that the biggest critique of neighborhood effects research is the simple fact that neighborhoods are themselves penetrated by a host of external forces and contexts. Even calls for analytic sociology are not terribly helpful, focused as they are on the relation of micro and macro (in this case neighborhood) levels, as opposed to higher order structures.

What is needed is a truly systemic approach that seeks to theorize and study empirically the "articulation" function of the local community vis-à-vis the larger social world – how organizations and social networks differentially connect local residents to the cross-cutting institutions that organize much of modern economic, political and social life (Janowitz 1975; Marwell 2007). Here the unit of analysis would be relations *across* neighborhoods – not merely as a function of geographical distance (e.g. ties in adjacent neighborhoods) but of the actual networks of connections that cross-cut neighborhood and even metropolitan boundaries – what Manuel Castells (1996) called the "space of flows." To use a simple example, if I move from community A to community B across town, I establish a connection between those places. It follows that the dominant structural flows of exchange between neighborhoods of different racial and economic status help us to understand better the reproduction of persistent urban inequality.

To illustrate the argument so far, consider the well-known "Coleman boat" (Coleman 1990: chapter 1) as applied to neighbourhood effects. How this might work theoretically and empirically is shown in Figures 11.2 and 11.3 with respect to a theory of crime rates and neighborhood measurement typology, respectively. Here I conceive of ecometrics as the science of measuring emergent and global characteristics

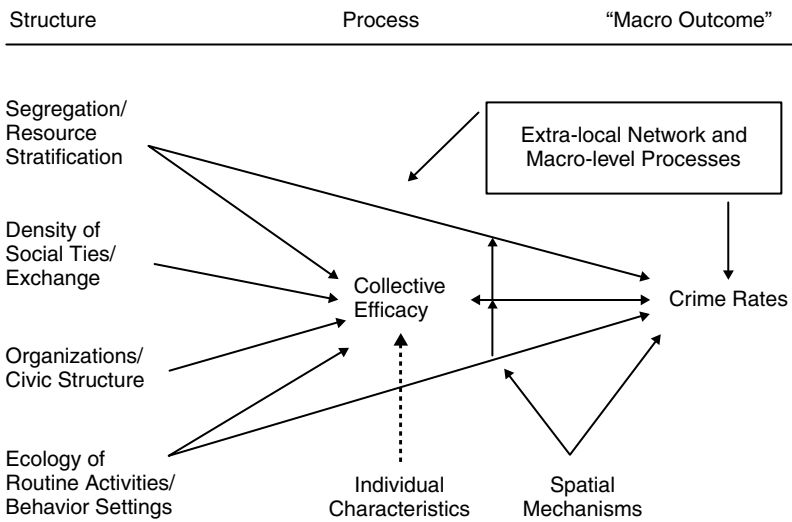


Figure 11.1 Neighborhood structure, social-spatial mechanisms, and crime rates.

of neighborhoods in a way that accounts for individual (selection) characteristics but at the same time goes beyond not just composition but also network ties and distributional properties such as inequality. It is important to note that traditional methods for establishing individual-level measurement properties are not the same as for neighborhood or structural level variations (Raudenbush and Sampson 1999). Ecometrically valid measures of collective efficacy, for example, tap between-area parameter variance based on informant ratings of others, not within-neighborhood or person-level variations. By the same token, however, ecometric procedures support the study of contextual effects on individuals, but once again in a way that broadens our scope to include more than just the compositional manifestations of the aggregation of individual properties about the self.

A series of recent studies have shown the independent association of collective efficacy with individual outcomes (e.g. depression, early sexual initiation, self-rated health, violence and individual efficacy in conflict avoidance) and macro associations across neighborhoods, especially rates of crime events and collective efficacy.[4] Note that Figure 11.1 portrays the importance of structural background (stratification),

[4] For a review of the evidence testing collective efficacy theory and crime see Pratt and Cullen (2005); for health see Browning and Cagney (2003).

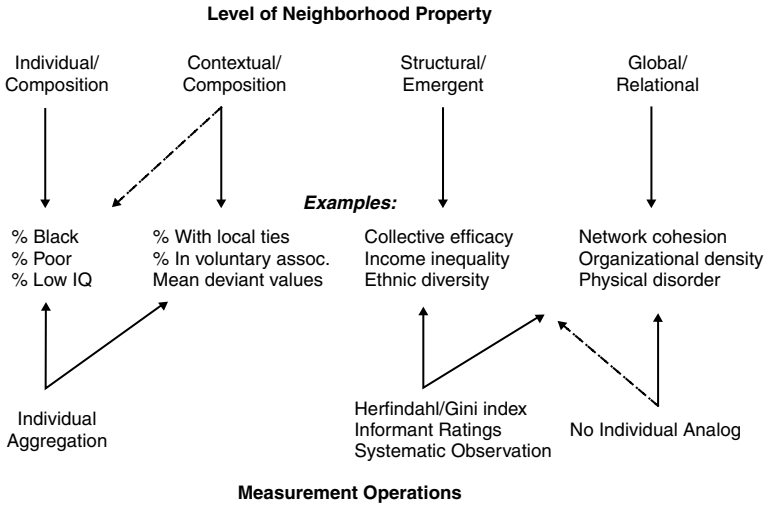


Figure 11.2 Ecometric typology of neighborhood properties and measurement.

organizations, the ecology of routines, and spatial dynamics (with both conditioning and direct effects), along with the hypothesized mediating role of collective efficacy (which can in turn be reciprocally influenced by crime and individual level factors). I do not claim this representation is exhaustive, and in ongoing work I am augmenting the model to include missing ingredients (e.g. networks of organizational ties). But for now I wish to focus on key hypotheses and assumptions about the role of collective efficacy as a process or mechanism accounting for crime rate variations.

The measurement typology in Figure 11.2 draws in part from the conceptual framework of Lazarsfeld and Menzel (1961). I define contextual or compositional characteristics as made up of the aggregation of individual characteristics, whether demographic (e.g. mean income) or an individual's ego-level connections (e.g. the mean level of having a friend in the neighborhood or mean voting).[5] *Structural* properties are

[5] Lazarsfeld and Menzel's typology described these as *analytical* properties, based on aggregated data about each member. I eschew the use of analytic here to avoid confusion with analytical sociology's emphasis on relational structures. I do make a distinction, however, by the theoretical intent of measurement: contextual measures such as mean friendship ties aim to describe a collective property (akin to a network) and not individual attributes, even though the measurement operation and result stem from the aggregation of individual characteristics.

based on data about the relations among group members or about the unit as a whole but may be derived from individual data. *Global* properties are not based on properties of individual members and cannot be derived solely from them (Lazarsfeld and Menzel 1961). Examples of a structural concept and measurement operation would include the collective efficacy of the group based on informant ratings or income inequality derived from the distribution of incomes. Global properties might include spatial proximity to a factory or the network cohesion of community leaders, where the latter measurement operations are based on the pattern of ties and not individuals. The dotted line is meant to indicate that while certain structural properties may not have a conceptual analog at the individual level (e.g. inequality), individual-level data go into their variable construction. Taken together, Figures 11.1 and 11.3 describe a theoretical and measurement approach that can be used to study neighborhood effects.

The conceptual orientation of Figure 11.1 is also meant to reflect both spatial proximity and cross-cutting linking processes that permeate the entire set of relationships. Such higher-order processes apply to what are typically considered intra- and inter-neighborhood factors. A case in point is when constraints such as neighborhood poverty are determined in part by government housing policies at the city, state and national level, or when social mechanisms such as collective efficacy are hypothesized to be influenced by the extra-local political and organizational resources that communities can bring to bear on local issues. The centralization of network ties among community leaders, for example, is related to community-level variations in trust independent of population composition. Trust appears to have both historical and cross-cutting network dimensions tied to organizational mechanisms (Sampson and Graif 2009).

Integrating the causal forces of history, cumulative advantages and higher-order networks is a major challenge to analytic sociology as currently conceived. In particular, the "supervenience" claim that a group (macro unit) has the properties it has by virtue of the lower order (individual) properties and relations of micro units (Hedström and Bearman 2009: 10) would seem to be overly restrictive. In the case at hand, I hypothesize that a key property at the macro level (say a neighborhood's capacity) is its position in the larger structure (e.g. spatial proximity to jobs, network position of the neighborhood, racial isolation) that can vary even though micro-level composition or relations internally are the same. How neighborhood processes are tied to the larger social, political, economic, historical and moral order of the city is thus a system-wide process that has few exemplars in extant research

and is not easily conceptualized within the confines of methodological or even structural individualism.

Individual selection and experiments reconsidered

So far I have been discussing how neighborhoods are studied "top down" (neighborhood effects on individuals), "side to side" (analysis of *between-neighborhood* rates of behavior), and from the "bird's-eye" or higher-order view of the city. But how neighborhoods change and how city dynamics are brought about must also simultaneously be considered from "from the bottom up." In particular, selection bias due to individual sorting is a near ubiquitous refrain in the burgeoning literature on neighborhood effects, whether at the pure macro level or "contextual level" in the Coleman scheme.[6] There is widespread concern that because individuals make choices and differentially sort themselves by place – non-randomly – estimates of neighborhood effects on individual outcomes are biased. By and large the response to this critique is to view selection as a statistical problem to be controlled away and not something of substantive interest in itself.

There is something odd about the way social scientists approach selection, however, almost as if they are spooked into thinking that choice renders the environment impotent. Scientifically, this approach is incorrect, just as a focus on genetics absent the environment is incorrect, or the flippant statement that people (the selectors) kill people, not guns, is incorrect as a full explanation. A common and increasingly influential way of thinking is that we must "randomize" with the use of experimental methods to address selection bias. This shift in research emphasis is perhaps not surprising given that experiments have long been cloaked in the mantle of science because of their grounding in the randomization paradigm, the putative cure for the ills of selection. If anything, the lure of experiments is increasing in the social sciences and in the field of neighborhood effects, where the belief that experiments are "a superior research strategy" for assessing causality is fast becoming a mantra (Oakes 2004: 1929).

I believe adopting this position is short-sighted, and have argued for alternative ways to conceive of neighborhood effects, selection bias, causal knowledge and, at bottom, the social structure of inequality. First of all, if we want to learn about the causal effects of neighborhood-

[6] Economists are often viewed as being at the forefront of the selection-bias concern but the sentiment is widely shared in the social sciences, with the seminal critique of neighborhood effects research offered by Jencks and Mayer (1990).

level processes in an experimental design, the most direct intervention is to randomly assign at the level of *neighborhoods* or other social units, not individuals (see Boruch and Foley 2000). Methodological or even "structural individualism" (Hedström and Bearman 2009: 8), it seems to me, has tended to brush past this important and scientifically legitimate question that originates at the macro level. One can imagine, for example, an experiment that randomly assigned neighborhoods to receive a manipulable "macro" treatment, such as community policing or mixed-income housing, that is hypothesized to promote collective efficacy. Because of random assignment, neighborhoods would be presumed equivalent on all "individual-level" selection factors, rendering the latter irrelevant as confounders. If collective efficacy were significantly influenced after the randomized intervention we may then speak of a *neighborhood-level causal effect*. Now, one cannot randomize collective efficacy itself in practice, but if further research in this strategy links collective efficacy to crime rates controlling for legitimate confounders, we then have a plausible set of evidence for collective efficacy as an explanatory mechanism linking the manipulable cause (e.g. community policing) to crime rates, absent an individual-level theory that connects the intervention to specific individuals committing acts of crime. Such neighborhood or population-level interventions are relatively more common in the public health arena (Sikkema *et al.* 2000), and from a public policy perspective may be more cost-effective than individual interventions. Regardless, my point is that we do not have to intervene at the level of an individual to demonstrate a causal effect linked to neighborhoods (see Figure 11.1). Our experimental research designs to date have been limited based on theoretical predilections biased toward methodological individualism and lack of imagination rather than pure scientific logic.

Second, as important as experiments might be in theory, a deep understanding of causality requires a theory of mechanisms no matter what the experiment (micro or macro) or statistical method employed: estimation techniques do not equal causal explanatory knowledge. Heckman (2005) has recently articulated what he calls the "Scientific Model of Causality," where the goal is to confront directly and achieve a basic understanding of the social processes that select individuals into causal "treatments" of interest. Relying on randomization via the experimental paradigm (even when logistically possible) sets aside the study of how mechanisms are constituted in a social world defined by the interplay of structure and purposeful choice.

Third, as noted above, individual perceptions and preferences are formed in part as a reaction *to the environment*, in this case neighborhood factors. And of course all behavior is situated in a context,

neighborhood being one. Consider, for example, the effect of neighborhood racial composition on perceptions of disorder, which triggers attributions (stigma) and actions (like "white flight" or abandonment of neighborhoods) which in turn reinforce initial beliefs or stereotypes and can lead to a cascade or dynamic process of decline (Sampson 2009). This process might be thought of as yet another kind of neighborhood effect, what we might call a "neighborhood perceptual mechanism" that sets in motion a causal chain of events. Controlling for such endogenous pathways, a typical approach, distorts the role of neighborhood context.

Social mechanisms and the reproduction of inequality

Patrick Sharkey and I recently examined selection into neighborhoods of varying types as an essential ingredient in the larger project of understanding neighborhood effects. We also examined the social *consequences* of residential selection (Sampson and Sharkey 2008). Our question became one of how individual mobility decisions combined to create spatial flows that define the ecological structure of inequality, an example of what Coleman (1990: 10) more broadly argued is a major under-analyzed phenomenon – micro-to-macro relations in.

In this section of the chapter I build on this work to describe briefly the sources and consequences of sorting for the reproduction of racial economic inequality in the lives of individuals, and for the reproduction of a stratified urban landscape at the macro level. My example comes from an integrated set of longitudinal, multi-level data – the Project on Human Development in Chicago Neighborhoods (PHDCN). The dispersion and movement of the sample was considerable – families moved all across Chicago, to the suburbs, and to many other states as well but were tracked no matter where they moved. From the parent interviews we geocoded detailed address information at all three waves of the survey which we in turn linked to census tract codes across the United States, allowing us to map changes in the neighborhood residential locations of sample members occurring over the course of the survey. This information was then merged with demographic and structural data on all census tracts in the United States.

We found that a number of previously unobserved factors that represent hypothesized sources of selection bias in studies of neighborhood effects were, despite the litany of suspicions raised in the literature, of surprisingly minimal importance in actual or revealed neighborhood selection decisions. Residential stratification falls powerfully along race/ethnic lines and socio-economic position, especially income and education. These are for the most part the only surviving

factors that explain a significant proportion of variance in neighborhood attainment. Even after controlling a comprehensive array of theoretically motivated covariates (some forty overall) that included previously unstudied aspects of locational attainment – such as depression, criminality and social support – the results were unchanged. Somewhat surprisingly, it follows that longitudinal studies accounting for neighborhood selection decisions and a fairly simple yet rigorous set of individual and family stratification measures may make for a reasonable test of neighborhood influences. This result alone is a testament to the power of context.

We next found that whites and Latinos living in neighborhoods with growing populations of non-whites were more likely to exit Chicago, providing evidence that realized mobility stems, in part, from a response to structural changes in the racial mix of the origin neighborhood. The same is not true of black families – the data suggested that it is not African-Americans' preference for same-race neighbors that seems to matter as much as whites' and Latinos' eagerness to exit neighborhoods with growing populations of blacks. Ironically, then, neighborhood conditions appear to matter a great deal for influencing neighborhood selection decisions, suggesting a unique kind of neighborhood effect – *sorting as a social process*.

An implication of our analysis is that the causes of residential moves may be less important than their aggregate consequences for the reproduction of a racialized hierarchy of place at the structural level. We pursued this line of inquiry in two ways by responding to Coleman's (1990: 10) call for the study of emergent micro–macro relations. We specifically explored the structural pattern of flows connecting origin neighborhoods in Chicago to destinations anywhere in the United States. Neighborhoods we classified based on location within or outside Chicago (mostly suburban Chicago), the dominant neighborhood racial/ethnic group, and poverty status – neighborhoods with median income in the bottom quartile of Chicago's distribution are defined as "poor." The pattern of flows renders visible the structural consequences of "selection." First that only tiny flows of Chicago residents produce upward mobility in the sense of crossing the boundaries of the racial and ethnic hierarchy that is present in Chicago neighborhoods and well beyond.[7] By far the most common outcome is to stay at

[7] See Sampson and Sharkey (2008: 24) for details. They depict transitions across neighborhoods undertaken by at least 5 percent of the residents in each origin neighborhood. Consistent with a decomposition of change approach, "stayer" mobility pathways we defined as residents moving addresses but remaining in the given neighborhood

one's original address, a crucial ingredient in reproducing the system of place stratification. The next most common transition leads movers into neighborhoods of the same subtype. Circulation within African-American neighborhoods is especially common, as seen in the 20 percent of families in segregated, poor black neighborhoods who change addresses but remain in the "ghetto"; similarly, 18 percent of residents in black non-poor neighborhoods move into different neighborhoods but of the same type. Twenty percent of families in black poor neighborhoods leave, but to other segregated black areas.

The dominant flows crossing a racial or ethnic boundary serve to reinforce the hierarchy of neighborhoods rather than undermine it:

1. Four out of the five dominant pathways out of Chicago are to white non-poor areas.
2. White-origin neighborhoods (all of which are non-poor) generate the largest pathway out of Chicago.
3. White neighborhoods in Chicago are a favored destination but they do not send to any other neighborhood subtype save one – Latino non-poor.

Five percent make the last transition, although this flow consists almost entirely of Latinos moving from predominantly white origin areas into Latino neighborhoods. Race-specific flows also document that 80 percent of whites transition into (or remain in) neighborhoods that are predominantly white *and* non-poor, whether inside or outside of the city. By contrast, when blacks and Latinos leave Chicago they continue to live in areas that are markedly less affluent than their white counterparts. Considering the improvements in neighborhood quality associated with mobility outside of the city, this pattern reveals one of the mechanisms by which whites maintain an ongoing advantaged position in the hierarchy of neighborhoods, even in a multi-ethnic and economically diverse area such as Chicago. Residential choice is highly structured.

Conclusion

In this chapter I have argued for a focus on neighborhood-level theory and the direct measurement of "ecometric" properties as a way to improve the

subtype over the course of the survey. By focusing on the most prominent transitions, the analysis was designed to complement analysis of individual trajectories by showing how aggregate movement connects neighborhoods and thereby produces a linked network of stratification.

study of neighborhood effects and the social mechanisms that produce them. I have articulated the concept of collective efficacy as a key neighborhood process implicated in the explanation of crime-rate variations (Figure 11.1). Unlike most research intentions on neighborhood effects to claim a hierarchical or "top down" primacy of neighborhoods over the individual, I have considered "side to side," "bottom up" and "bird's eye" orientations along with issues of measurement (Figure 11.2), social causality, methodological individualism, extra-local spatial processes, perceptions as causation, and selection bias all as a way to address what I consider a modified analytic sociology "Coleman project". My argument implies that there are many ways to conceive of neighborhood effects, with the dominant "top down" pathway important but far from the only link worth considering. Even when "contextual effects" are is at issue, its conceptualization is typically one where mediating pathways are inappropriately controlled (the ubiquitous attempt to render "all else equal"). But if neighborhood conditions influence cognitive perceptions and evaluations, they are causally implicated in individual behavior, residential selection and social structure alike. Social life is emergent and neighborhoods may be conceived as a complex social system (Sawyer 2005).

Selection bias has emerged as a stated threat to this view of neighborhood-effects research and there has been a surge of calls for more individual-level studies embedded within experimental designs. I have argued that this move does not satisfy an "emergentist" or social mechanisms perspective. The results summarized here support instead the notion that neighborhood selection is inextricably a social process that unfolds within an ordered, yet constantly changing, residential landscape. In examining the sources and social consequences of residential sorting, neighborhood selection is thus not an individual-level confounder or a "nuisance" that arises independent of social context (Bruch and Mare 2006; Heckman 2005). Rather, selection and sorting are part of a dynamic social process of neighborhood stratification that reproduces racially shaped economic hierarchies and that leads to an apparently durable equilibrium absent macro-level interventions (Sampson and Sharkey 2008). This leads me to what in some quarters would be considered a radical conclusion: *sorting is at once a causal mechanism and a social process*.

From this view, it follows that just because environments are selected does not mean that they lack causal import. One can distinguish conceptually between why individuals are exposed to certain neighborhood environments (selection) and what the causal effect is on their actions of the influences to which they are subjected (Wikström 2006: 88). The first does not negate the second and both are part of neighborhood

effects. In addition, selection does not mean that we must demote the role of observational data in the causal study of social mechanisms associated with neighborhood effects. The experimental impulse that is now hegemonic is reductionist in practice, estimating parameters that while unbiased and of policy relevance are not necessarily those of most theoretical interest at the macro level (Sampson 2008). Namely, conceptualizing a sequence of independent "treatments" may induce clarity of thought with regard to estimating a single unbiased parameter, but such a move does not describe the social world, much less explain it. Theory about interdependent social mechanisms and observational data are still necessary.

Ultimately, then, higher-order processes that induce structure at the neighborhood level require a different way of thinking than the individualist and largely micro-level approaches of existing experimental paradigms. To tackle causal processes that take on historical and institutional dimensions that range over long periods of time, sometimes decades, requires a more flexible conception of causality than that offered by individual experiments and even, dare I say, methodological individualism (Salmon 1998). Causal processes may be mediated by human agents, but like gene–environment interactions, the notion of a "main effect" at the gene (individual) level is scientifically misleading. Put differently, to say that the micro level takes ontological precedence, as some forms of analytic sociology claim it does, seems parallel to the idea of a main effect of genes, whereby the environment "supervenes" genes. But supervenience (Hedström and Bearman 2009: 10–11) cannot explain the *interaction* between genes and environment (Lewontin 2000), just as variation at the contextual level cannot be reduced to individual actions or the micro base – if there is interaction no one level is privileged.

Despite the real promise of analytical sociology, I conclude that methodological individualism would do well to grant social context and macro-level factors an equal forum in theories of neighborhood effects.[8] We need to measure heretofore unmeasured social level constructs and theorize the role of neighborhood context in influencing

[8] See also the pragmatist critique of methodological individualism in Gross (2009). In my view Gross' pragmatist philosophy has some appeal, but his tendencies to "contingent" and culturally specific strictures render the approach problematic for those interested in general theory (as opposed to grand theory) and empirical generalizations. Rather than presume a cultural or "particularized" account is necessary a priori, I would look instead to the theoretical question and whether the empirical problem to be solved demands such an account. In this spirit, I would also note Coleman's (1990: 5) pragmatic rejection of approaches that insist causal explanations must always be taken down to the individual level: "The explanation is satisfactory if it is useful for the particular kinds of intervention for which it is intended."

perceptions and interactions that take on emergent properties, in turn reinforcing and shaping rates of behavior, ultimately forming a key part of the explanation for how social structures emerge and change.

REFERENCES

Bandura, Albert. 1997. *Self Efficacy: The Exercise of Control*. New York: W.H. Freeman.

Boruch, Robert and Ellen Foley. 2000. "The honesty experimental society: sites and other entities as the units of allocation and analysis in randomized trials," in L. Bickman (ed.), *Validity and Social Experimentation: Donald T. Campbell's Legacy*. Thousand Oaks: Sage.

Browning, Christopher R. and Kathleen A. Cagney. 2003. "Moving beyond poverty: neighborhood structure, social processes and health," *Journal of Health and Social Behavior* 44: 552–71.

Brubaker, Roger. 1996. *Nationalism Reframed: Nationhood and the National Question in the New Europe*. Cambridge and New York: Cambridge University Press.

Bruch, Elizabeth E. and Robert D. Mare. 2006. "Neighborhood choice and neighborhood change," *American Journal of Sociology* 112: 667–709.

Castells, Manuel. 1996. *The Rise of the Network Society*. Oxford: Blackwell.

Coleman, James S. 1988. "Social capital in the creation of human capital," *American Journal of Sociology* 94: S95–120.

1990. *Foundations of Social Theory*. Cambridge, MA: Harvard University Press.

Firey, Walter. 1947. *Land Use in Central Boston*. Cambridge, MA: Harvard University Press.

Grannis, Rick. 1998. "The importance of trivial streets: residential streets and residential segregation," *American Journal of Sociology* 103: 1530–64.

Gross, Neil. 2009. "A pragmatist theory of social mechanisms," *American Sociological Review* 74: 358–79.

Heckman, James J. 2005. "The scientific model of causality," *Sociological Methodology* 35: 1–97.

Hedström, Peter. 2005. *Dissecting the Social: On the Principles of Analytical Sociology*. Cambridge University Press.

Hedström, Peter and Peter Bearman. 2009. "What is analytic sociology all about? An introductory essay," in Peter Hedström and Peter Bearman (eds.), *The Oxford Handbook of Analytical Sociology*. Oxford University Press.

Hunter, Albert. 1974. *Symbolic Communities: the Persistence and Change of Chicago's Local Communities*. University of Chicago Press.

Janowitz, Morris. 1975. "Sociological theory and social control," *American Journal of Sociology* 81: 82–108.

Jencks, Christopher and Susan E. Mayer. 1990. "The social consequences of growing up in a poor neighborhood," in Lawrence Lynn and Michael McGreary (eds.), *Inner-City Poverty in the United States*. Washington, DC: National Academy Press, 111–86.

Lazarsfeld, Paul F. and Herbert Menzel. 1961. "On the relation between individual and collective properties," in Amatai Etzioni (ed.), *Complex Organizations*. New York: Holt, Rinehart & Winston, 422–40.

Leventhal, Tama and Jeanne Brooks-Gunn. 2000. "The neighborhoods they live in: the effects of neighborhood residence on child and adolescent outcomes," *Psychological Bulletin* 126: 309–37.

Lewontin, Richard. 2000. *The Triple Helix: Gene, Organism, and Environment*. Cambridge, MA: Harvard University Press.

McRoberts, Omar. 2003. *Streets of Glory: Church and Community in a Black Urban Neighborhood*. University of Chicago Press.

Marwell, Nicole. 2007. *Bargaining for Brooklyn: Community Organizations in the Entrepreneurial City*. University of Chicago Press.

Morenoff, Jeffrey D., Robert J. Sampson and Stephen Raudenbush. 2001. "Neighborhood inequality, collective efficacy, and the spatial dynamics of urban violence," *Criminology* 39: 517–60.

Oakes, J. Michael. 2004. "The (mis)estimation of neighborhood effects on causal inference for a practicable social epidemiology," *Social Science and Medicine* 58: 1929–52.

Portes, Alejandro and Julia Sensenbrenner. 1993. "Embeddedness and immigration: notes on the social determinants of economic action," *American Journal of Sociology* 98: 1320–50.

Pratt, Travis and Frances Cullen. 2005. "Assessing macro-level predictors and theories of crime: a meta-analysis," in Michael Tonry (ed.), *Crime and Justice: A Review of Research*. University of Chicago Press, 373–450.

Raudenbush, Stephen W. and Robert J. Sampson. 1999. "'Ecometrics': toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods," *Sociological Methodology* 29: 1–41.

Salmon, Wesley C. 1998. *Causality and Explanation*. New York: Oxford University Press.

Sampson, Robert J. 1988. "Local friendship ties and community attachment in mass society: a multi-level systemic model." *American Sociological Review* 53: 766–79.

2008. "Moving to inequality: neighborhood effects and experiments meet social structure," *American Journal of Sociology* 114: 189–231.

2009. "Disparity and diversity in the contemporary city: social (dis)order revisited," *British Journal of Sociology* 60: 1–31.

2011. *Neighborhood Effects: Social Structure and Community Processes in the American City*. University of Chicago Press (forthcoming).

Sampson, Robert J. and Corina Graif. 2009. "Neighborhood networks and processes of trust," in Karen S. Cook, Margaret Levi and Russell Hardin (eds.), *Whom Can We Trust? How Groups, Networks, and Institutions Make Trust Possible*. New York: Russell Sage Foundation, 182–215.

Sampson, Robert J. and Patrick Sharkey. 2008. "Neighborhood selection and the social reproduction of concentrated racial inequality." *Demography* 45: 1–29.

Sampson, Robert J., Jeffrey D. Morenoff and Felton Earls. 1999. "Beyond social capital: spatial dynamics of collective efficacy for children," *American Sociological Review* 64(5): 633–60.

Sampson, Robert J., Jeffrey D. Morenoff and Thomas Gannon-Rowley. 2002. "Assessing 'neighborhood effects': social processes and new directions in research." *Annual Review of Sociology* 28: 443–78.

Sampson, Robert J., Jeffrey D. Morenoff and Stephen W. Raudenbush. 2005. "Social anatomy of racial and ethnic disparities in violence." *American Journal of Public Health* 95: 224–32.

Sampson, Robert J., Stephen W. Raudenbush and Felton Earls. 1997. "Neighborhoods and violent crime: a multilevel study of collective efficacy," *Science* 277: 918–24.

Sawyer, R. Keith. 2005. *Social Emergence: Societies as Complex Systems*. New York: Cambridge University Press.

Searle, John R. 1995. *The Construction of Social Reality*. New York: The Free Press.

Sikkema, Kathleen, J.A. Kelly, R.A. Winett, L.J. Solomon, V.A. Cargill, R.A. Roffman, T.L. McAuliffe, T.G. Heckman, E.A. Anderson, D.A. Wagstaff, A.D. Norman, M.J. Perry, D.A. Crumble and M.B. Mercer. 2000. "Outcomes of a randomized community-level HIV prevention intervention for women living in 18 low-income housing developments," *American Journal of Public Health* 90: 57–63.

Small, Mario. 2009. *Unanticipated Gains: Origins of Network Inequality in Everyday Life*. New York: Oxford University Press.

Sørensen, Aage B. 1998. "Theoretical mechanisms and the empirical study of social processes," in P. Hedström and R. Swedberg (eds.), *Social Mechanisms: An Analytical Approach to Social Theory*. Cambridge University Press.

Suttles, Gerald D. 1972. "The defended community," in Gerald D. Suttles (ed.), *The Social Construction of Communities*. University of Chicago Press, 21–43.

Tilly, Charles. 1973. "Do communities act?" *Sociological Inquiry* 43: 209–40.

Warren, Donald. 1975. *Black Neighborhoods: an Assessment of Community Power*. Ann Arbor: University of Michigan Press.

Wikström, Per-Olof. 2006. "Individuals, settings, and acts of crime: situational mechanisms and the explanation of crime," in Per-Olof Wikström and Robert J. Sampson (eds.), *The Explanation of Crime: Context, Mechanisms, and Development*. Cambridge and New York: Cambridge University Press.

2010. "Situational action theory," in Francis Cullen and Pamela Wilcox (eds.), *Encyclopedia of Criminological Theory*. Thousand Oaks: Sage.

Wikström, Per-Olof and Robert J. Sampson. 2003. "Social mechanisms of community influences on crime and pathways in criminality," in Ben Lahey, Terrie Moffitt and Avshalom Caspi (eds.), *Causes of Conduct Disorder and Serious Juvenile Delinquency*. New York: Guilford Press, 118–48.

2006. *The Explanation of Crime: Context, Mechanisms, and Development*. Cambridge and New York: Cambridge University Press.

Wikström, Per-Olof H., Vania Ceccato, Beth Hardie and Kyle Treiber. 2010. "Activity fields and the dynamics of crime: advancing knowledge about the role of the environment in crime causation," *Journal of Quantitative Criminology* 26: 55–87.

Wilson, William Julius. 1987. *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy*. University of Chicago Press.
Woodward, James. 2003. *Making Things Happen: a Theory of Causal Explanation*. Oxford University Press.

12 Social mechanisms and generative explanations: computational models with double agents

Michael W. Macy with Damon Centola, Andreas Flache, Arnout van de Rijt and Robb Willer

Traditionally, sociologists have tried to understand social life as a structured system of institutions and norms that shape individual behavior from the top down. In contrast, a new breed of social modelers suspect that much of social life emerges from the bottom up, more like improvisational jazz than a symphony orchestra. People do not simply play parts written by elites and directed by managers. We make up our parts on the fly. But if everyone is flying by the seat of their pants, how is social order possible? The puzzle is compounded by scale. Coordination and cooperation are relatively easy to attain in a jazz quartet, but imagine an ensemble with millions of musicians, each paying attention only to those in their immediate vicinity. Without a Leviathan holding the baton, what prevents the population from descending into a Hobbesian cacophony of all against all?

Social life from the bottom up

New and compelling answers to this question are being uncovered by social theorists using an innovative modeling tool developed in computer science and applied with impressive success in disciplines ranging from biology to physics – agent-based computational (ABC) modeling.

ABC models are *agent-based* because they take as a theoretical starting point a model of the autonomous yet interdependent individual units (the "agents") that constitute a dynamic system. The models are *computational* because the individual agents and their behavioral rules

This chapter builds on and extends Flache and Macy (2009), Macy and Willer (2002), Centola and Macy (2005) and Macy and van de Rijt (2006) and Van de Rijt, A., D. Siegel and M. Macy (2009). Co-authors are listed alphabetically. The authors wish to thank the National Science Foundation (SBR 0241657 and SES-0432917) and the Netherlands Organization for Scientific Research (NWO-VIDI-452–04–351) for support during the time that this research was conducted.

are formally represented and encoded in a computer program such that the dynamics of the model can be deduced using step-by-step computation from given starting conditions.

ABC modeling was originally developed in computer science and artificial intelligence as a technology to solve complex information processing problems based on autonomous software units. These units can each perform their own computations and have their own local knowledge, but they exchange information with each other and react to input from other agents. ABC models have been used to understand how spontaneous coordination can arise in domains as diverse as computer networks, bird flocks and chemical oscillators.

Increasingly, social scientists are using this same methodology to better understand the complexities of social life as well. Despite their technical origin, agents are inherently social. Agents have both a cognitive and a social architecture (Gilbert and Troitzsch 1999; Wooldridge and Jennings 1995). Cognitively, agents are heuristic and adaptive. Socially, agents are autonomous, interdependent, heterogeneous and relational. *Heuristic* means that agents follow simple behavioral rules, not unlike those that guide much of human behavior, such as habits, rituals, routines, norms, and the like (Simon 1982). *Adaptive* means that actions have consequences that alter the probability that the action will recur, as agents respond to feedback from their environment through learning and evolution. *Autonomous* agents have control over their own goals, behaviors and internal states and take initiative to change aspects of their environment in order to attain those goals. They interact with little or no central authority or direction. However, autonomy is constrained by behavioral and strategic *interdependence*. Behavioral interdependence means agents influence others in response to the influence that they receive, through persuasion, sanctioning and imitation. Strategic interdependence means the consequences of each agent's decisions depend in part on the choices of others. *Heterogeneity* relaxes the assumption in system dynamics models (and many game theoretic models as well) that populations are composed of representative agents.

Finally, and perhaps most importantly, agents are relational; that is, they interact locally with "neighbors," such that population dynamics are an emergent property of local interaction. ABC models allow researchers to study how local decision heuristics and network topology interact to produce highly complex and often surprising global patterns.

Can social scientists learn something from models developed for understanding computer networks, bird flocks or chemical oscillators? Agent modelers believe they can, for several reasons. First, ABC models show how very simple rules of local interaction can generate

highly complex population dynamics that would be extremely difficult (if not impossible) to model using traditional methods. Second, these models show how Durkheimian "social facts" can emerge *sui generis* at the population level, even when these properties do not exist at the level of the interacting agents. Third, these models can be used as virtual laboratories, to reveal the micro mechanisms responsible for highly complex social phenomena.

The microfoundations of social complexity

Like game theory, ABC modeling is a formal implementation of "methodological individualism," the search for the microfoundations of social life in the actions of intentional agents. This is the major difference between ABC modeling and an earlier generation of equation-based methods of computer simulation such as system dynamics. Traditional computer simulation modeled population behavior as a unified system. These models generated population patterns based on interactions among system-level processes, usually expressed as rates of change among system components. ABC models generate population patterns as an emergent property of local interactions among autonomous decision-makers. Instead of a model of the population, we have a population of interdependent models, each corresponding to an individual actor. Contrary to conventional intuitions, ABC models have repeatedly demonstrated how the emergence of macro-social patterns out of micro-social interactions is not always a matter of simple aggregation.

ABC modeling also differs from game theory in relaxing the behavioral and structural assumptions required for mathematical tractability. The attraction to ABC modeling centers on the ability to use computation to solve problems that would otherwise be mathematically intractable, such as systems with nonlinear dynamics, complex networks and sensitive dependence on initial conditions. As in the physical and life sciences, social scientists have begun to appreciate that the complexity of social systems cannot always be understood using traditional deductive methods.

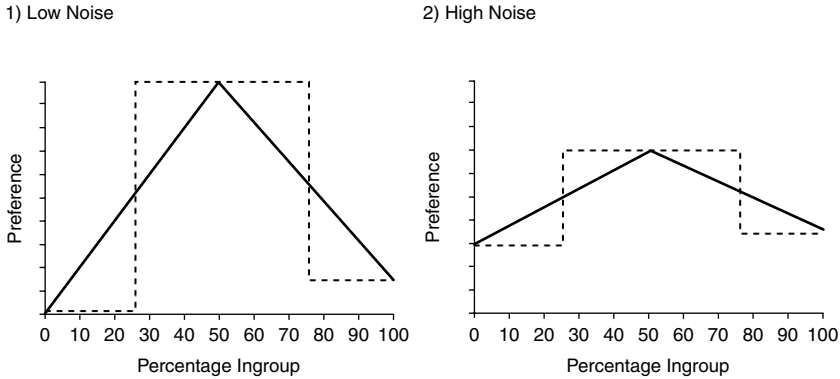
Thomas Schelling's (1971) canonical tipping model of residential segregation was among the first efforts to extend game theory by letting the players walk through the problem rather than trying to solve the problem mathematically. Consider a city that is highly segregated, such that every resident has neighbors with identical cultural markers (such as ethnicity). If the aggregate pattern were assumed to reflect the attitudes of the residents, one might conclude from the ethnic distribution that the population was highly parochial and intolerant of diversity. Yet Schelling's tipping model shows that this need not be the case. He assumes that

residents are tolerant of ethnic diversity but do not want to become a minority in the neighborhood. The model shows that perfectly segregated neighborhoods will form even in a population that is tolerant of diversity. The aggregate pattern should therefore not be taken to reflect the underlying attitudes of the constituent individuals. The first agent who moves out to avoid being in the minority makes the imbalance even worse for the agent's co-ethnic neighbors. This ethnic flight, in turn, adds another majority-group member to the receiving neighborhood, thereby making the ethnic distribution even less tolerable for the minority members.

A computational extension of the Schelling model

Schelling's classic model did not require a computer. The model was implemented on a large checkerboard using red and blue poker chips. This implementation made it necessary to keep the model extremely simple. As a consequence, the model imposed three simplifying assumptions that can be relaxed by implementing the model on a computer. First, Schelling assumed that agents are tolerant of diversity but nevertheless prefer co-ethnic neighbors. It is not all that surprising that segregation would emerge in a population in which everyone is ethnocentric, even if they tolerate diversity. We do not need computer simulation to immediately recognize that total segregation is an equilibrium in this population. Suppose instead that agents have multiculturalist preferences for diversity, such that they would choose a mixed neighborhood over one that was entirely co-ethnic. Would this prevent the population from tipping into segregation? Second, Schelling assumed that decisions about where to live were deterministic. Suppose there is a small amount of noise in decision-making. Might this destabilize a segregated equilibrium? Third, Schelling assumed that agents were indifferent to ethnic composition on either side of a single threshold. Suppose the probability of moving increases continuously with the discrepancy between the observed and desired ethnic composition of the neighborhood. Put differently, suppose the agents are highly sensitive to even very small changes in the ethnic composition of their neighborhood.

Van de Rijt *et al.* (2009) relaxed these three simplifying assumptions by implementing the preference function used in a mathematical model developed by Zhang (2004; see also Pancs and Vriend 2007), in which agents strongly prefer neighborhoods with a mixed composition that reflects the population distribution of ethnicity. In addition, agents' moves are stochastic, based on a probability distribution that corresponds to their preferences. Agents are thus least likely to move out when the current neighborhood reflects the ethnic diversity of the population and most likely to move out when the local population is homogenous,



The discontinuous line indicates a discontinuous (threshold) preference for diversity. The solid line indicates a linear preference. These preferences are weaker relative to chance in panel 2.

Figure 12.1 Multicultural preferences.

but slightly less so if this homogeneity favors their own ethnic group. Finally, in one condition, agents are sensitive to small changes in ethnic composition, and in the other condition, agents have a threshold level of ethnic homogeneity above which they move out in search of a neighborhood with greater diversity, as depicted in Figure 12.1.

When the noise level was set to zero and the preference function was linear, the results confirmed Zhang's (2004) mathematical results – the population segregated. Moreover, this outcome was highly robust to the introduction of a small amount of noise in the decision function. However, when the linear function was replaced with a threshold function, the population remained integrated. A multiculturalist population with a threshold preference for diversity is less likely to tip into segregation compared to a similarly multiculturalist population that is sensitive to small changes in the ethnic composition of their neighborhood.[1]

The mechanism that allows a multicultural population to segregate is the asymmetry (evident in Figure 12.1) in dissatisfaction with neighborhoods that are ethnically imbalanced. This dissatisfaction is slightly

[1] Using a computational model in which agents prefer as many co-ethnic neighbors as possible, Bruch and Mare (2006) reached the opposite conclusion – that tipping at the individual level (i.e. a threshold for moving out) leads to tipping at the population level (i.e. a cascade leading to segregation). However, van de Rijt *et al.* (2009) showed that this result was caused by an error in their model, and that the integration they observed with linear preferences was possible only when ethnic preferences were sufficiently weak relative to chance, an error Bruch and Mare acknowledged (2009) in response.

weaker when the imbalance favors the ingroup than when an identical imbalance favors the outgroup. Although this difference in dissatisfaction is very small compared to the much larger difference in the preference for balanced vs. imbalanced neighborhoods, the small ingroup bias is sufficient to tip the population into a highly segregated pattern when agents are sensitive to small changes in ethnic composition. Even in a population that is highly integrated, there will be some neighborhoods that will not be perfectly balanced – one of the two groups will be slightly favored. In these minimally imbalanced neighborhoods, agents will have a very low probability of moving out, but the probability will nevertheless be slightly higher if the imbalance favors the outgroup. Over time, this small bias inevitably causes the level of imbalance to increase, which in turn confronts agents with an ever-starker choice between neighborhoods that favor – by an ever-widening margin – one group over the other. The only possible outcome is then a maximally segregated population that leaves everyone dissatisfied except the lucky few who manage to end up on the borders between the ghettos.

However, agents with threshold preferences manage to remain integrated despite a comparable asymmetry in their dissatisfaction with imbalanced neighborhoods. That is because the asymmetry only applies to agents who live in neighborhoods that exceed their tolerance level for ethnic imbalance. As long as the level of imbalance remains tolerable, these agents do not care whether the imbalance favors their own group or the other. And as long as the agents do not care, the level of imbalance remains within their tolerance level.

In short, these results show that Schelling actually understated the possibility for segregation to emerge in a population that does not seek segregation. He showed how this can happen in a population that tolerates ethnic diversity up to a point. We now see how segregation can emerge even in a population that prefers integration, so long as the residents are sufficiently sensitive to small changes in ethnic composition.

The double edge of networks

An ABC model of cooperation among friends also showed how individual behaviors can aggregate with opposite effects on the behavior of the population. In network analysis, an "edge" is an undirected tie between two nodes, such as two friends or two neighbors. Flache and Macy (1996) discovered that friendship networks can have a "double edge," meaning that the tie between two friends can have opposite effects on the solidarity of a group to which both friends belong. On the one side, people usually care more about approval from their friends

than from strangers, and this can promote cooperative behavior that is enforced informally by peer pressure to conform to group obligations, particularly in groups whose members are highly dependent on one another for social approval (such as residents of a small town with limited opportunities for outside friendships). It can be proven mathematically that it is an equilibrium when everyone contributes and everyone approves of their friends.[2] Moreover, the greater the dependence on friends for approval, the stronger the effect of peer pressure in reinforcing this equilibrium. Yet when Flache and Macy tested the effects of dependence on the group using an ABC model of contribution to a public good, the result was surprising: the more dependent people were on approval from their friends, the less willing they were to cooperate. What went wrong? Although dependence on friends for social approval made informal sanctions more effective, it also made friends reluctant to risk friendship by criticizing one another. The mechanism is structural. Because friendship is dyadic, coordination on mutual approval is far more likely than the multilateral coordination required for the exchange of contribution for approval. As a consequence, agents exchanged approval directly for approval, even when both friends failed to contribute to the group.

Contrary to intuition, a stronger need for approval at the micro level did not lead to more effective social control at the macro level.

An agent-based model of naked emperors

Centola *et al.* (2005) discovered another population dynamic with surprising microfoundations. They modeled the curious tendency for people to pretend to believe something they know not to be true, and to disparage those who disagree, a behavioral pattern that is portrayed in Hans Christian Andersen's story of the *Emperor's New Clothes*. Everyday examples include college students who celebrate and encourage intoxication through drinking games, yet who privately express discomfort with excessive consumption (Prentice and Miller 1993). The foible is not limited to students. We all know "naked" scholars whose incomprehensible writings are applauded by those who pretend to understand and appreciate every word, and who disparage skeptics for being intellectually shallow. The pattern extends to closet deviants who publicly disparage others so as to avoid suspicion, such as anxious citizens in totalitarian regimes who denounce their neighbors so as to affirm their loyalty to the regime (Kuran 1995).

[2] For a comparison of computational and mathematical solutions, see Flache *et al.* (2000).

Centola *et al*. wanted to know if a cascade of compliance could trap a population into thinking a norm was highly popular, when in fact almost everyone shared a strong (but private) wish that the norm would go away. In addition, they examined the structural conditions that might make a population highly vulnerable to such a cascade. Their model assumes a population of agents who are willing to comply with and enforce a norm they privately oppose, but only if the pressure to comply is sufficiently strong. These agents live in a clustered network, but with some overlap between the clusters. The model is initialized with a population that contains only a tiny fraction of vigilant "true believers" who truly support the norm and will always sanction deviant neighbors. The rest of the population opposes the norm, and in the absence of enforcement, the skeptics refuse to comply.

The model demonstrated that widespread enforcement of an unpopular norm can be an equilibrium in a population where the fear of exposure as an imposter is sufficient to motivate enforcement as a way to signal the sincerity of compliance. Yet this equilibrium cannot be reached from a random start in a fully connected population. However, if interaction is restricted to the local neighborhood, and these neighborhoods sufficiently overlap, the equilibrium can be reached, even with only a tiny fraction of "true believers."

The computational model also identified a surprising condition that is necessary for these cascades to succeed. Contrary to an extensive literature on small worlds (Watts and Strogatz 1998) and "the strength of weak ties" (Granovetter 1973), long-range ties that bridge across clusters actually inhibit the propagation of unpopular norms compared to networks of equal size and density but with fewer ties between clusters. Why this sensitivity to network topology? "Shortcuts" across the network allowed agents in conformist neighborhoods to see deviants in other neighborhoods, which in turn reinforced their private opposition to the norm.

These results deepen our understanding of the dynamics of a puzzling social phenomenon that has perplexed social scientists for years. Further, the results motivated a follow-up laboratory experiment designed to study the false enforcement of an unpopular norm (Willer *et al*. 2009). In the face of social pressure to conform, participants in the experiment not only agreed to a belief they knew to be false, but also publicly penalized a deviant who stated what the participants privately knew to be true. Yet when sanctioning was private, participants favored the deviant. In a related paper, Centola and Macy (2007) investigated in greater detail why shortcuts across a network can inhibit rather than facilitate the propagation of a social contagion, such as conformity to an emergent but unpopular norm. A series of computational experiments highlighted the importance

of a qualitative distinction between "simple contagions" with an "infection" threshold of one (such as disease or information) and "complex contagions" with any threshold higher than one (such as the willingness to act on information). Social contagions are typically complex, due to the social and material costs faced by early adopters of risky innovations, controversial practices, or new technologies whose benefits require widespread usage (e.g. email). As a result, the probability of adoption increases with the proportion of one's neighbors who have already adopted. This points to the need for an important caveat for Granovetter's (1973) theory of the "strength of weak ties" that bridge between network clusters. Although clustered sources of information tend to be redundant, clustered sources of advice, persuasion or legitimation are essential for the spread of many social contagions.

The virtues and vices of ABC modeling: the problem of double agents

Agent-based simulations are often criticized on three counts: artificiality, triviality and reliability:

1. Computational models, such as mathematical models, tell us about a highly stylized and abstract world, not about the world in which we live. In particular, agent-based models rely on highly simplified models of rule-based human behavior that neglect the complexity of human behavior.
2. Computational models cannot tell us anything that we do not already know, given the assumptions built into the model and the inability of a computer to do anything other than execute the instructions given to it by the program.
3. Unlike mathematical models, simulations are numerical and therefore cannot establish lawful regularities or generalizations.

1. *Realism*. The defense against these criticisms centers on the principles of complexity and emergence. In complex systems, very simple rules of interaction can produce highly complex global patterns. Although this principle has been mainly applied to physical and biological systems, Simon (1998: 53) draws out the striking implication for the study of human behavior – the possibility that "the apparent complexity of our behavior is largely a reflection of the complexity of the environment," including the complexity of interactions among strategically interdependent, adaptive agents. The simulation of artificial worlds allows us to explore the complexity of the social environment by removing the cognitive complexity (and idiosyncrasy) of constituent individuals. "Human beings," Simon contends, "viewed as behaving systems, are

quite simple." We follow rules, in the form of norms, conventions, protocols, moral and social habits, and heuristics. Although the rules may be quite simple, they can produce global patterns that may not be at all obvious and are very difficult to understand.

In short, the artificiality of agent-based models is a virtue, not a vice. Agent-based modeling does not necessarily "aim to provide an accurate representation of a particular empirical application" (Axelrod 1997: 5). Rather, the goal "is to enrich our understanding of fundamental processes that may appear in a variety of applications." When computational models are used for simulation or to make predictions or for training personnel (e.g. flight simulators), the assumptions need to be highly realistic, which usually means they will also be highly complicated (Axelrod 1997: 5). "But if the goal is to deepen our understanding of some fundamental process," Axelrod continues, "then simplicity of the assumptions is important and realistic representation of all the details of a particular setting is not."

There are several advantages of artificiality. First, lawful regularities may be obscured by conditions in the empirical world that block or distort their expression. Holland (1995: 146) offers a classic example: Aristotle's mistaken conclusion that all bodies come to rest, based on observation of a world that happens to be infested with friction. Had these observations been made in a frictionless world, Aristotle would have come to the same conclusion reached by Newton and formulated as the principle that bodies in motion persist in that motion unless perturbed. Newton avoided Aristotle's error by studying what happens in an artificial world in which friction had been assumed away. Ironically, "Aristotle's model, though closer to everyday observations, clouded studies of the natural world for almost two millennia" (Holland 1995: 146). In contrast, Newton's model generates more accurate predictions than does Aristotle's, once the effects of friction are reconsidered. More generally, it is often necessary to step back from the world we experience, into one that is more abstract, in order to see the empirical world more clearly.

For example, Schelling's neighborhood-segregation model made no effort to be "realistic" or to resemble an observed city. Rather, the "residents" live in a highly abstract cellular world without traffic, poverty, crime, lousy schools, train tracks, zoning laws, red-lining, housing prices, etc. This artificial world shows that all neighborhoods have an underlying tendency toward segregation even when residents are tolerant of diversity, so long as residents do not want to become a local minority. This hypothesis can then be tested in observed neighborhoods. Of course, real world observations may not confirm the model predictions, but this does not detract from the value of the experiment. On the contrary, empirical disparities lead the investigator to look for

conditions that were not present in the model that might account for the discrepancy with the observed outcome.

In short, agent-based models "have a role similar to mathematical theory, shearing away detail and illuminating crucial features in a rigorous context" (Holland 1995: 100). Nevertheless, Holland reminds us that we must be careful. Unlike laboratory experiments, computational "thought experiments" are not constrained by physical reality and thus "can be as fanciful as desired or accidentally permitted. Caution and insight are the watchwords if the computer-based model is to be helpful" (Holland 1995: 96).

2. *The results are wired in*. The second criticism centers on the inability of computers to act in any way other than how they were programmed to behave. This criticism overlooks the need to "determine the logical consistency of a set of propositions and the extent to which theoretical conclusions actually follow from the underlying assumptions" (Prietula *et al.* 1998: xv). Indeed, a properly designed computational experiment can reveal highly counter-intuitive and surprising results. This theoretical possibility corresponds to the principle of *emergence* in complex systems. Biological examples of emergence include life that emerges from non-living organic compounds and intelligence and consciousness that emerge from dense networks of very simple switch-like neurons. Schelling's (1971) neighborhood model provides a compelling example from social life: segregation is an emergent property of social interaction that does not necessarily reflect the simple aggregation of individual preferences. And Centola *et al.* (2005) show that norms that enjoy widespread public support do not necessarily reflect the distribution of private beliefs. In short, ABC models can be extremely useful in identifying surprising macro-level implications of a set of micro-level assumptions. That these implications are "wired in" does not in any way detract from the value of the computational experiments. On the contrary, by revealing unexpected population patterns that are "wired in" to a set of behavioral and structural assumptions, ABC models demonstrate the danger of trusting "intuition" and "common sense" instead of formal modeling.

Nevertheless, the problem of wiring in the results can also be very serious: when the computational results are sensitively dependent on hidden assumptions that are theoretically arbitrary. Sensitivity to theoretically motivated assumptions is not a problem – on the contrary, that is the reason for running computational experiments in which these assumptions are carefully manipulated. However, results that are an artifact of hidden assumptions that do not correspond to the theory that motivated the research are completely worthless and misleading. This criticism points to the need to make all assumptions explicit and

to test the robustness of the results to alternative specification of those assumptions that are theoretically arbitrary.

3. *Computational models are not deductive*. A third criticism is that computational models, unlike mathematical models, are numerical, not deductive. This has two important implications. First, unlike deductive models that do not depend on particular numerical inputs, ABC models generate population patterns that vary across the parameter space for reasons that can be very difficult to identify, especially if the models are overly complicated. Agent models can reveal correlations between inputs and outputs even if the modeler has no idea why these results are being observed. Although computer programs must be as logically tight as the proof of a theorem, the program logic is processual rather than causal. Hence, we can know with certainty that a set of implications follow from a set of assumptions, without knowing why. Without explicit identification of the causal mechanisms, we cannot rule out the possibility that the results are nothing more than artifacts of particular modeling technologies or even bugs in the source code.

This is the dilemma of "double agents" posed by tremendous advances in the speed and power of inexpensive desktop computers. Computation allows formal models of methodological individualism to become much richer than the comparatively stylized mathematical models in game theory. However, this computational power has a double edge. ABC models can become so complicated that it is no longer possible to identify the causal mechanisms responsible for the results. Computational agents make it possible to identify the logical implications of a set of behavioral and structural assumptions that could not be mathematically derived. Yet these same agents can also obscure the underlying causal mechanisms by opening the door to rich theoretical elaboration.

The solution is the development of a standardized methodology of systematic theoretical elaboration based on step-by-step experimentation. The fundamental principle underlying this methodology is Einstein's imperative to make models as simple as possible and no simpler than necessary. Modelers must resist the temptation to add bells and whistles that make the model more realistic or that satisfy an insatiable curiosity to find out what might happen if some new parameter were added. Nothing is gained by making a model more realistic if this also makes it impossible to identify the causal mechanisms underlying the correlations between inputs and outputs.

The second implication of this criticism is that computational results lack the generality of deductive conclusions (Holland 1995: 100). For this reason, computational models should not be used when a mathematical solution is available. However, mathematical models pay a high

price for the ability to generate deductive conclusions: they require simplifying assumptions in order to be mathematically tractable. While computational models risk becoming too complex to be understood, mathematical models risk becoming too simple to address important theoretical questions.

For example, Centola *et al.* (2005) show that universal enforcement of an unpopular norm can be an equilibrium, and that the structure of the network determines whether this equilibrium can be reached from a starting point that is far from equilibrium. It can be proven mathematically that if everyone enforces the obligation to enforce, then no one has an incentive to deviate unilaterally. However, this does not imply that this equilibrium will ever be reached. That question might also be answered mathematically for a randomly mixed or fully connected population. The problem is that few social networks are random and large populations are almost never fully connected. If we want to understand how an enforcement cascade might spread successfully on a complex network, then computational methods are likely to be required.

More generally, ABC models lack the generality of deductive conclusions, but they may nevertheless be the only practical method to capture the out-of-equilibrium dynamics that guide a population of agents from some initial condition to a stable social arrangement or from one equilibrium (when perturbed) to another. Agent-based models also extend equilibrium analysis to cases in which social stability arises even though individual strategies are constantly changing. A serious limitation of Nash equilibrium – the principal solution concept in analytical game theory – is the assumption that population stability is predicated upon the stability of the individual members, that is, the unwillingness of anyone to change strategies unilaterally. Young (1993, 1998) has proposed an alternative mathematical solution – stochastic stability – that overcomes this limitation of static equilibrium analysis. However, it remains to be shown that stochastic stability is mathematically tractable for games that are played in complex networks. Computational methods remain the most practical method for modeling dynamic equilibrium as well as out-of-equilibrium population behavior, especially for populations with complex network structures.

A dynamic equilibrium obtains when the forces pushing the system in one direction are precisely balanced by countervailing forces, as in models of system dymamics (Forrester 1968). ABC models, however, do not theorize these forces at the system level. Rather, the forces emerge out of the interactions of large numbers of interdependent but autonomous agents. For example, van de Rijt *et al.* (2009) show that segregation can be an equilibrium in a population of unhappy multiculturalists. The

agents are moving at every opportunity, but the population level distribution remains homeostatic.

ABC models have another important benefit – the ability to model heterogeneity among the actors. It may be mathematically impossible to derive equilibria for a population of players with different preferences, different resources, and incomplete information about the preferences and resources of other players. Because ABC models are implemented at the individual level, they can be applied to large populations of heterogeneous agents just as easily as to populations in which everyone is exactly like everyone else and knows everything that everyone else knows.

In sum, mathematical models have an important advantage over ABC models in deriving conclusions deductively rather than by numerical computation. Whenever mathematically feasible, deductive methods should be preferred, e.g. in identifying Nash equilibria. However, deductive methods have limited application to out-of-equilibrium dynamics, especially for populations that are not randomly or fully connected. ABC models facilitate exploration of the effects of network topology on aggregate dynamics, such as the study of complex contagions by Centola and Macy (2007). Computation also invites rich elaboration of models that can make the models much more realistic than the highly stylized mathematical models used in game theory. This is not necessarily an advantage. Excessive complexity can obscure identification of the causal mechanisms that underlie correlations between parameter inputs and the behavior of the population. It may therefore be useful to use rich models as testbeds to discover regularities that can then be derived mathematically using a stripped down version of the model, as van de Rijt *et al*. did in their (2009) study of the segregative effects of linear and discontinuous ethnic preferences. Despite the inability to yield deductive generalizations, ABC models remain essential for the study of complex adaptive systems, whether social, physical or biological. As Holland concludes (1995: 195), "I do not think we will understand morphogenesis, or the emergence of organizations like Adam Smith's pin factory, or the richness of interactions in a tropical forest, without the help of such models."

REFERENCES

Axelrod, Robert. 1997. *The Complexity of Cooperation*. Princeton University Press.

Bruch, E. and R. Mare. 2006. "Neighborhood choice and neighborhood change," *American Journal of Sociology* 112: 667–709.

2009. "Preferences and pathways to segregation: reply to van de Rijt, Siegel, and Macy," *American Journal of Sociology* 114: 1181–98.

Centola, D and M. Macy. 2005. "Social life in silico: the science of artificial societies," in Susan Wheelan (ed.), *Handbook of Group Research and Practice*. Thousand Oaks: Sage Publications, 273–83.

2007. "Complex contagions and the weakness of long ties," *American Journal of Sociology* 113: 702–34.

Centola, D., R. Willer and M. Macy. 2005. "The emperor's dilemma: a computational model of self-enforcing norms," *American Journal of Sociology* 110: 1009–40.

Flache, A. and M. Macy. 1996. "The weakness of strong ties: collective action failure in a highly cohesive group," *Journal of Mathematical Sociology* 21: 3–28.

2009. "Social dynamics from the bottom up: agent-based models of social interaction," in P. Bearman and P. Hedström (eds.), *Oxford Handbook of Analytical Sociology*. Oxford University Press.

Flache, A., M. Macy and W. Raub. 2000. "Do company towns solve free rider problems? A sensitivity analysis of a rational-choice explanation," in W. Raub and J. Weesie (eds.), *The Management of Durable Relations: Theoretical and Empirical Models for Households and Organizations*. Amsterdam: Thela Thesis.

Forrester, J. 1968. *Principles of Systems*. Waltham: Pegasus Communications.

Gilbert, N. and K. Troitzsch. 1999. *Simulation for the Social Scientist*. Milton Keynes: Open University Press.

Granovetter, M. 1973. "The strength of weak ties," *American Journal of Sociology* 78: 1360–80.

Holland, J. 1995. *Hidden Order: How Adaptation Builds Complexity*. Reading, MA: Perseus.

Kuran, T. 1995. *Private Truths, Public Lies: the Social Consequences of Preference Falsification*. Cambridge, MA: Harvard University Press.

Macy, M. and A. van de Rijt. 2006. "Game theory," in B. Turner (ed.), *The Cambridge Dictionary of Sociology*. Cambridge University Press.

Macy, M. and R. Willer. 2002. "From factors to actors: computational sociology and agent-based modelling," *Annual Review of Sociology* 28: 143–66.

Pancs, R. and N. Vriend. 2007. "Schelling's spatial proximity model of segregation revisited," *Journal of Public Economics* 91: 1–24.

Prentice, D. and D. Miller. 1993. "Pluralistic ignorance and alcohol use on campus: some consequences of misperceiving the social norm," *Journal of Personality and Social Psychology* 64: 243–56.

Prietula, M., K. Carley and L. Gasser. 1998. *Simulating Organizations: Computational Models of Institutions and Groups*. Cambridge, MA: MIT Press.

Schelling, T. 1971. "Dynamic models of segregation," *Journal of Mathematical Sociology* 1: 143–86.

Simon, H. 1982. *Models of Bounded Rationality*. Cambridge, MA: MIT Press.

1998. *The Sciences of the Artificial*. Cambridge, MA: MIT Press.

Van de Rijt, A., D. Siegel and M. Macy. 2009. "Neighborhood chance and neighborhood change: a comment on Bruch and Mare," *American Journal of Sociology* 114: 1166–80.

Watts, D. and S. Strogatz. 1998. "Collective dynamics of 'small-world' networks," *Nature* 393: 409–10.
Willer, R., K. Kuwabara and M. Macy. 2009. "The false enforcement of unpopular norms," *American Journal of Sociology* 115: 451–90.
Wooldridge, M. and N. Jennings. 1995. "Intelligent agents: theory and practice," *Knowledge Engineering Review* 10: 115–52.
Young, H.P. 1993. "The evolution of conventions," *Econometrica* 61: 57–84.
1998. *Individual Strategy and Social Structure: an Evolutionary Theory of Institutions*. Princeton University Press.
Zhang, J. 2004. "Residential segregation in an all-integrationist world," *Journal of Economic Behavior and Organization* 54: 533–50.

13 Relative deprivation *in silico*: agent-based models and causality in analytical sociology

Gianluca Manzo

Introduction

The concept of relative deprivation is one of the most frequently used notions in economics (see Clark *et al.* 2008), in social psychology (see Tyler *et al.* 1997: ch. 2; Walker and Smith 2001: ch. 1) and in sociology (see Cherkaoui 2001; Coleman 1990: ch. 8; Lundquist 2008). Despite its diffusion, formal analyses of the mechanisms generating rates and feelings of relative deprivation are far less common.[1]

In sociology, the most notable exceptions are, on the one hand, Boudon's (1982: ch. 5; 1979: 52–6) analysis – later taken up by Kosaka (1986) and Yamaguchi (1998) – and, on the other, Burt's (1982: ch. 5, 191–8) contribution.

These analyses, however, have a different focus. The first group have the following characteristics:

1. They are interested in the rates of relative deprivation.
2. They tend to demonstrate that the relation between objective opportunity structure and proportion of dissatisfied actors can be both negative and positive.
3. They implicitly refer to actors who compare themselves with a given group as a whole (global comparisons).

By contrast, Burt's model can be characterized as follows:

1. It focuses on the individual feelings of deprivation.

I wish to express my gratitude to Andrew Abbott, Carlo Barone, Thomas Fararo and Kenji Kosaka for reading and commenting on a first draft of this paper and to Amy Jacobs and Barbara Cowell for correcting and revising my English.

[1] Davis (1959) seems to be the first attempt to formalize the ideas at the heart of *The American Soldier*. His model is concerned with the proportion of deprived actors and it supposes completely socially unstructured comparisons between actors. The model, however, does not contain any generative mechanism of the rate of relative deprivation.

2. It does not address the question about the positive or negative nature of the relation between objective opportunity structure and the intensity of feelings of dissatisfaction.
3. It takes into account comparisons between people who are embedded in social networks (local comparisons).

My aim here is to develop a unified theoretical framework which enables us to analyze formally the relation between these different aspects at the same time. In particular, I will try to demonstrate two statements:

1. The four-way relation between the attractiveness of the goods at stake, the opportunity structure, the percentage of dissatisfied actors, and the intensity of their feelings of dissatisfaction may take a variety of forms except the most sought-after one, i.e. the "*more opportunities, less dissatisfied-and-less-intensely-dissatisfied actors*" pattern.
2. The presence of dyadic interactions can significantly modify certain aspects of this four-way relation such as it originally appears in a microcosm whose actors are entirely isolated and where only global comparisons are made.

Compared with the above-mentioned formal analyses, an additional distinctive trait here is that I have sought to solve these problems by programming and studying an agent-based model (Ferber 1999; Gilbert 2007; Miller and Page 2007).[2]

In the context of this book, this application serves a second purpose. The chapter is intended as an illustration of the potentialities of agent-based modeling as a methodological support for the two main aspects of the conception of causality analytical sociology is built on, i.e. "generativity" and "counterfactuality."[3]

According to the first criterion, causal claims rest on the possibility to demonstrate that the relation between two happenings ultimately comes from an underlying bundle of structured triads "entities/

[2] This computational method has recently been singled out for its conceptual flexibility and computational power in economics (Axtell 2000; Epstein 2006; Tesfatsion and Judd 2006), finance (Mathieu *et al*. 2005), political science (Axelrod 1997; Cederman 2001), geography (Sanders 2007) and at least partially in sociology (Hummon and Fararo 1995; Macy and Flache 2009; Macy and Willer 2002; Sawyer 2003).

[3] Agent-based methodology has recently been put on the analytical sociology agenda (Hedström 2005: ch. 6; Hedström and Bearman 2009). I discuss this link more deeply elsewhere (see Manzo 2007a, 2007b, 2010). Let us also notice that the elective affinities between generative epistemology and agent-based methodology have already been pointed out (Cederman 2005; Epstein 2006: chs. 1–2). By contrast, to my knowledge, no explicit bridge has yet been built between the counterfactual account of causality and agent-based models.

properties/activities," that is to say a "mechanism" (see Machamer *et al*. 2000: 3). Such a conception of causality was first outlined by Harré (1972: 115–19, 136–7), who called it "generative theory of causality"; it then progressively spread in statistics (Cox 1992), economics (Simon 1979) and sociology (Boudon 1979; Fararo 1989, 2009; Goldthorpe 2001; Hedström 2004). Analytical sociology is programmatically building on this idea (see Hedström 2005: ch. 2; Hedström and Bearman 2009; Hedström and Swedberg 1998: 7).

On the other hand, the counterfactual account of causality basically states that the causal character of the relation between two happenings ultimately rests on the possibility to demonstrate that if, say, *X* had not occurred, *Y* would not have occurred. Deeply grounded in philosophy (see Lewis 1973 and, more recently, Woodward 2000), such a conception of causality has widely been accepted in economics earlier than in sociology (see Morgan and Winship 2007; Winship and Morgan 1999). As some recent contributions suggest (see Hedström and Udéhn 2009; Hedström and Ylikoski 2010), counterfactuals are also entering epistemological agenda of analytical sociology.

From a methodological point of view, "generativity" and "counterfactuality" raise two different, although related, problems. The generative criterion requires to demonstrate that a given set of loops between structures, behaviors and interactions produces the aggregate patterns of interest. An agent-based model allows to provide this demonstration. Its internal structure allows the design of multi-level artificial mechanisms while its dynamic makes it possible to transform the mechanism in a process, which is what one is looking for when one wants to determine what a mechanism is able to bring about.

On the other hand, the counterfactual criterion demands to evaluate the degree to which a given alteration of the mechanism at hand modifies the aggregate patterns of interest. Agent-based models allow to easily perform this task. When one explores the parameter space and the internal structure of an agent-based model, one is indeed studying the sensitivity of the outcomes to the mechanisms and its initial conditions. In this sense, this computational technique provides a powerful tool to create and analyze "potential outcomes" *in silico*.

It is worth noticing, however, that to claim that agent-based models represent flexible causally generative and counterfactual devices does not amount to state they are able to produce empirically validated causal statements on their own. In this regard, it is important to distinguish two steps:

1. the analysis of how posited mechanisms work and the high-level patterns they generate; and

2. the empirical validation of them.

The first task requires to construct microcosms which run in accordance with one or another set of rules capable of generating one or another set of individual and collective states. Here agent-based models are useful and necessary. By contrast, the second task is not specific to analytic sociology: it only requires injecting empirical information at the entrance to or exit from an agent-based model. We already have a broad spectrum of tools (qualitative and quantitative) for doing this.

To solve the problem of discovering real-world causal relations, we obviously have to integrate the two phases. But to claim that we should test a mechanism empirically before submitting it to rigorous formal study is to reverse the order in which the problems should be solved.[4]

The chapter is organized as follows. I first give an overview of the literature on relative deprivation and I posit some useful conceptual distinctions. I then present the theoretical structure of the agent-based model I built in order to study the rate and feelings of relative deprivation at the same time. Lastly, I discuss the computational results obtained by simulating this artificial society under several parameter settings. The conclusion summarizes the questions I have addressed as well as the main results and limitations of the analysis.

A useful analytical distinction: RD frequency and RD intensity

The empirical observations which gave rise to sociological literature on relative deprivation (hereafter noted RD) all noted an inverse relation between actors' perceptions of the conditions they act in and the "objective" quality of those conditions.[5]

Stouffer and his colleagues (1965 [1949]: vol. I, pp. 52, 125) were the first to use the concept explicitly to explain this seemingly paradoxical

[4] I tried to satisfy both requirements in my analyses of educational inequalities in France and in Italy (see Manzo 2009a). Two other good examples of sociological empirically calibrated agent-based models are Hedström (2005: ch. 6) and Bruch and Mare (2006).

[5] The most well-known is certainly the inverse correlation at the core of *The American Soldier* (Stouffer *et al*. 1965 [1949]: 251–2) between promotion rates in the army and subjective perception of opportunities for promotion. But, before *The American Soldier*, Tocqueville (1955 [1856]: bk. III, ch. 4, p. 176) had observed that "it was precisely in those parts of France where there had been the most improvement that popular discontent ran highest." Durkheim (1951 [1897]: bk. II, ch. V, p. 244) noted that "an unusual increase in the number of suicides is observed with this collective renaissance." After *The American Soldier*, Runciman (1966: 3) acknowledged that "dissatisfaction with the system of privileges and rewards in a society is never felt in an even proportion to the degree of inequality to which its various members are subject."

correlation. The hypothesis implicit in this "interpretative intervening variable," as Merton (1957: 229) described it, is that actors' assessments of their objective opportunities actually depend on their standards of comparison (Stouffer *et al*. 1965 [1949]: vol. I, p. 125).[6]

While empirical observation of a linear inverse relation between opportunity structure and people's perceptions of those opportunities was what first motivated the use of the RD concept, the problem of how general that relation is has not yet been completely resolved.[7]

This problem is complex because it arises from two distinct but overlapping dimensions. On the one hand, the RD phenomenon involves two aspects; on the other hand, a large variety of mechanisms responsible for them can be at work.

On the first point, one should carefully distinguish between RD frequency – i.e. the proportion of actors who do not have what they want – and RD intensity: the strength of the feeling actors associate with this discrepancy (see Runciman 1966: 10; see also Elster 2007: 58). This suggests that the mechanisms that move a certain number of actors to perceive a discrepancy between reality and their desires may be different from those that engender their specific reactions to this assessment. From this in turn it follows that the relations between conditions of well-being and subjective perceptions of those conditions can take different forms depending on which aspect of RD is being studied and the type of mechanisms mobilized.[8]

[6] Runciman (1966: 10) was the first to give a more developed definition: "We can roughly say that A is relatively deprived of X when (i) he does not have X, (ii) he sees some other person or persons, who may include himself at some previous or expected time, as having X (whether or not this is or will be in fact the case), (iii) he wants X, and (iv) he sees it as feasible that he should have X." A pioneering definition developed in social psychology adds a fifth component: "lack[s] a sense of responsibility for failure to possess X" (Crosby 1976: Table 1).

[7] The authors of *The American Soldier* themselves seemed aware of the problem: "To be conservative, we should limit our conclusion by saying that a force with relatively less promotion chances tended to have a larger proportion of men speaking very favorably of promotion opportunities than a force with greater promotion chances" (Stouffer *et al*. 1965 [1949]: 257). The point was mentioned in passing by Merton (1957: 237, n. 7): "presumably, the relationship is curvilinear, and this requires the sociologists to work out toward the conditions under which the observed linear relation fails to obtain." Runciman (1966: 19–20) took up the point nearly ten years later: "this relation is both complicated and variable … it can as well take the form of an inverse correlation as a direct one" (1966 247). In economics, the Easterlin paradox holding that "raising the incomes of all does not increase the happiness of all" (Easterlin 1973: 4) has been repeatedly analyzed (cf. Clark *et al*. 2008) to demonstrate that a positive relation between income and satisfaction with life does exist, not only at the individual level but also at the aggregate level and not only within a given country but also among countries (Wolfers and Stevenson 2008).

[8] This analytic distinction appears clearly in contemporary social psychology definitions of RD: "a judgment that one is worse off compared to some standard; this judgment

On the second point, RD generative mechanisms can be inscribed in a basic analytic space using axes that correspond to the comparison reference points that actors choose (for a more specific analytical map, see Gambetta 1998: 114–19). Two main types are usually considered in social psychology (Tyler *et al.* 1997: ch. 2):

1. actor-specific reference points, namely one's own past condition or expectations (intrapersonal comparisons); and
2. reference points external to the actor, namely other individuals or groups (inter-individual and intergroup comparisons).

Recent studies have attempted to show that these two types of comparisons actually proceed from a single, more general type known as counterfactual comparisons: "comparisons of one's current outcomes with outcomes that one might have obtained but did not" (Olson and Roese 2002: 266).[9]

Compared with the statistical analysis of observational data (see Clark *et al.* 2008: 111–15), constructing formal models of RD-generating mechanisms and analyzing them deductively seems an attractive way of trying to establish what the form of the relation between opportunity structure and RD frequency/intensity is. This strategy indeed enables us first to establish all the outcomes logically associated with a given mechanism (or several), and, then, to locate, within this range of possibilities, the section of the real world covered by the empirical data under study.

As I said, Boudon's formal model suggested that the relation between opportunity and individual satisfaction can be both negative and

is linked, in turn, to feelings of anger and resentment" (Tyler *et al.* 1997: 17); "a subjective state that shapes emotions and cognitions and influences behavior" (Pettigrew 2002: 353).

[9] From an historical point of view, we find intrapersonal comparisons in Tocqueville's *Old Regime and the French Revolution* (1955 [1857]: 177): "Dazzled by the prospect of a felicity undreamed of hitherto and now within their grasp, people were blind to the very real improvement that had taken place and eager to precipitate events"; "Patiently endured so long as it seems beyond redress, a grievance comes to appear intolerable once the possibility of removing it crosses men's mind" (see Cherkaoui (2005: ch. 1) for a perceptive reading of these mechanisms in Tocqueville's work). We also find intrapersonal comparisons, with astounding parallelism, in Durkheim's thinking: "Thus, the more one has, the more one wants, since satisfactions received only stimulate instead of filling needs" (Durkheim 1951 [1897]: 248); and "The less limited one feels, the more intolerable all limitation feels" (Durkheim 1951 [1897]: 254). Durkheim also seems to have been sensitive to intergroup comparisons: "Lack of power, compelling moderation, accustoms men to it, while nothing excites envy if no one has superfluity" (Durkheim 1951 [1897]: 254). The social comparison reference points at the heart of *The American Soldier* (Stouffer *et al.* 1965: 251) prompted Merton's analysis (1957: chs. VII and VIII) of the concept of "reference group." Runciman (1966: 24–5) combined the two, positing a loop between a rise in individual expectations and a rise in actor reference group level.

positive, a point that has been confirmed by Kosaka's and Yamaguci's re-analyses of the model. The mechanism which generates this result is simple: a combined set of rules, individual reasoning and interdependence structure that lead a certain number of actors to rationally hope to obtain more than they could objectively obtain (according to Gurr's (1970: 51) typology, this is a "aspirational deprivation" mechanism). In terms of the above distinctions between RD frequency and RD intensity, however, all these authors were only concerned with RD frequency. But what is the form of the relation between opportunity structure and RD intensity? Moreover, how are these three elements linked to one another and how is this threefold relation modified when actors are embedded in some sort of relational structures?

An agent-based model of RD frequency and RD intensity

To answer these questions, I programed an agent-based model which contains six components. While the first five simply generalize Boudon's original model, the sixth one introduces a new module which quantifies the disappointment, envy and regret which dissatisfied actors may feel when intrapersonal comparisons, population-based or neighborhood-based inter-individual comparisons and counterfactual reasoning are at work.[10]

1. *Agents' opportunity structure*. It is specified by the following elements: (a) a population of N agents; (b) a limited numbers of two types of goods, G^1 and G^2; (c) the sum of G^1 plus G^2 is always equal to N; (d) G^1 and G^2 differ in attractiveness in the sense that the benefit B^1 (> 0) associated with G^1 is higher than the benefit B^2 (≥ 0) associated with G^2; (e) G^1 and G^2 also differ in accessibility in the sense that G^1 can only be obtained if agent spends C^1 (> 0 and $< B^1$) whereas G^2 can only be obtained if agent spends C^2 (≥ 0, $\leq B^2$ and $< C^1$); (f) all agents have enough resources to be able to spend C^1 or C^2.
2. *Agents' beliefs*. They build on the following elements: (a) each agent knows the number of G^1 and G^2 available in society but does not know the number of agents $A(S^1)$ and $A(S^2)$ who will respectively adopt strategy S^1 (spending C^1 to obtain B^1) or strategy S^2 (spending C^2 to obtain B^2); (b) each agent must therefore estimate the gain expected from S^1 ($G[S^1]$) compared to the gain expected from S^2

[10] Constructing and analyzing a multi-agent system is still a fairly costly operation (see Janssen *et al*. 2008). Here I used the agent-based simulation platform option (Railsback *et al*. 2006), specifically NetLogo 4.0.3 (Tisue and Wilensky 2004a, 2004b; Wilensky 1999).

($G[S^2]$) as a function of the number of agents $A(S^1)$ likely to opt for S^1 (and therefore the number likely to opt for S^2).[11]

3. *Agents' desires*. Each agent wishes to obtain a net benefit from his choice so that, for each number of agents $A(S^1)$ likely to opt for S^1, he will choose S^1 if and only if $G(S^1) - G(S^2) > r$ (where r is the minimum gain demanded).
4. *Agents' final choice*. Given the vector of choices for or against S^1 produced by the dynamic conjunction of an agent's beliefs and desire, the probability of the agent ultimately deciding for or against S^1 increases non-linearly as a function of the proportion of cases in which agent chooses S^1. Specifically, I chose a logistic function discretized by 10-unit intervals.[12]
5. *Agents' final gain*. Once agents have made their definitive choice for S^1 or S^2, G^1 and G^2 available in the system can be allocated to them. There are three possible situations:

   (a) If the number of agents who definitively opted for S^1 is exactly equal to the number of G^1, all agents are satisfied: those who wanted G^1 got G^1; the others, who wanted G^2, got G^2.
   (b) If the number of agents who definitively opted for S^1 is greater than the number of G^1, some of the agents who spent C^1 to obtain B^1 will actually only be able to obtain G^2. As there are no individual or social screening traits in this artificial world, agents receiving the lesser benefit at the higher cost are determined by random selection.
   (c) If the number of agents who ultimately opt for S^1 is below the number of G^1, the number of agents opting for S^2 will be above the number of available G^2. Given that the game rule stipulating

[11] In particular, as long as $A(S^1) < G^1$,

$$G[S^1] = B^1 - C^1 \quad [1]$$
$$G[S^2] = ((B^2 - C^2) * G^2) / A(S^2) \quad [2]$$

Instead, when $A(S^1) > G^1$,

$$G[S^1] = ((B^1 - C^1) * G^1) / A(S^1) + ((B^2 - C^1) * (A(S^1) - G^1)) / A(S^1)$$
$$= ((B^1 - B^2) * G^1 / A(S^1)) + (B^2 - C^1) \quad [3]$$
$$G[S^2] = B^2 - C^2 \quad [4]$$

It is worth noticing that neither Boudon (1982: 117) nor Kosaka (1986: 36–40) considered the case where $A(S^1) < G^1$. This omission is probably due to the fact that both authors were studying the model only for $B^2 = C^2 = 0$ and $r = 0$. Under the condition $r = 0$, the case of $A(S^1) < L^1$ is not of much interest because S^1 will always be more advantageous than S^2. But if the intention is to run the simulation on a vast range of parameter combinations, this generalization of agents' belief updating process has to be included.

[12] The two situations originally studied by Boudon (1979: 52–6) – S^1 is chosen in 50 percent of cases and S^1 is chosen in 100 percent of cases (S^1 as a dominant strategy) – thus

that B^1 cannot be obtained by spending only C^2 precludes allocating G^1 to these agents, the simplest solution is to randomly allot a zero-gain to the surplus of agents desiring G^2.

Thus programed, Boudon's original model can now generate not one but two types of RD (hereafter indicated as RD^1 and RD^2), whose frequency can be studied (hereafter respectively indicated RD^1_{freq} and RD^2_{freq}). In particular, RD^1 affects agents who, having chosen G^1, only got G^2 because there were not enough G^1 lots. By contrast, RD^2 affects agents who wanted G^2 but in fact got nothing given the rules of the game and because there were not enough G^2 lots.

6. *Agents' emotions*. The experience of RD^1 may generate a different bundle of feelings of dissatisfaction than the one generated by the experience of RD^2. In this connection, I posit that:

   (a) Intrapersonal comparisons will be made in both cases. The strength of the disappointment they generate is proportional to the size of the difference between expected gain and gain ultimately obtained.
   (b) Inter-individual comparisons also exist for both RD^1 and RD^2. The strength of the envy they generate is understood to be inversely proportional to the number of those who did not get what they wanted (i.e. RD^1 freq and RD^2 freq).[13]
   (c) RD^2 can also imply a specific source of dissatisfaction. Agents finding themselves in this situation may reason counterfactually as follows: "If the rules of the game were different, there wouldn't be a waste of G^1." They may think that non-allotted G^1 could be put back in the game at a lower price – exceptionally. My assumption here is that criticism of this kind, implicitly aimed at the rule system in effect, may give rise to regrets, and that the breadth of those regrets would be proportional to the number of non-allotted G^1 lots.[14]

become respectively the equilibrium point and the upper limit of a more general choice function. This choice then represents here the main source of heterogeneity among agents – a point Yamaguchi (1998) greatly insisted on. The sources of heterogeneity will be much more extensive in the sixth component of the model.

[13] According to Elster (1999: 141), envy is one of the most frequent comparison-based emotions (the ones "triggered by favorable or unfavorable comparisons with individuals with whom we will never interact"). More specifically, I am quantifying here the strength of this emotion by a mechanism implicitly postulated by Stouffer *et al.* (1965: 251), where the intensity of individual feelings of dissatisfaction is inversely related to the diffusion of failure.

[14] Elster (1999: 241–2) establishes a direct link between counterfactual reasoning and emotions: "Fifth, there are counterfactual emotions generated by thoughts of what

With $RD^1_{intensity}$ and $RD^2_{intensity}$ indicating the intensity of the dissatisfaction feeling perceived by agents experiencing respectively RD^1 and RD^2, we have the following simple representation of these three hypotheses:

$$RD^1_{intensity} = \alpha[(B^1 - C^1) - (B^2 - C^1)] + \beta(1/ RD^1_{freq}) \quad [5]$$

$$RD^2_{intensity} = \gamma[(B^2 - C^2) - (0 - C^2)] + \delta(1/ RD^2_{freq}) + \lambda[\text{non-allotted } G^1] \quad [6]$$

where $\alpha, \beta, \gamma, \delta$ and λ are random values drawn from uniform distributions [0, 0.5] that represent the idea that the three feeling-of-deprivation generative mechanisms operate differently from one individual to another.[15]

In truth, this formalization implies an additional supposition. The second term in [5] and [6] actually inversely links the intensity of envy felt by agents with the overall proportion of agents who find themselves in the same deprivation situation.

But we could reasonably allow that when agents are determining how strongly they think they have been penalized in not getting what they want compared to those who spent as much as they did and *did* get the desired lot, they only take into account local diffusion of RD^1 and RD^2. While this hypothesis seems reasonable – one point in its favor is that it does not require us to suppose that agents have permanent knowledge of the overall state of the system – it also raises the problem of defining what is "local."

As indicated by the second term of [7] and [8], my hypothesis here is that what makes up the horizon within which agents assess the diffusion of RD situations is the set of dyadic ties they are embedded in ("neigh" refers to agent's "neighborhood"; i.e. the agents he is in direct contact with).[16]

$$RD^1_{intensity} = \alpha[(B^1 - C^1) - (B^2 - C^1)] + \beta(1/ RD^1_{freq}[^{neigh}]) \quad [7]$$

$$RD^2_{intensity} = \gamma[(B^2 - C^2) - (0 - C^2)] + \delta(1/ RD^2_{freq}[^{neigh}]) + \lambda[\text{non-allotted } G^1] \quad [8]$$

might have happened but didn't – regret, rejoicing, disappointment, elation – and wistful subjective emotions generated by thoughts of what might still happen, albeit with insufficient probability to generate hope or fear."

[15] To obtain $RD^1_{intensity}$ and $RD^2_{intensity}$ values that vary between two given extremes, we can standardize each term of [5] and [6] (see below note 21). It would also be useful to study how the model behaves if we substitute a "ratio" (or a "log-ratio") for the difference in the first term of [5], [6], [7] and [8], since the algebraic properties of this functional form are considerable (see Jasso 2008). Finally, notice that the first term of equations [5] and [6] can be simplified, respectively, to (B1–B2) and to (B2), so expressing the idea that the strength of the disappointment generated by intrapersonal comparisons is supposed to be proportional to the size of the expected benefit that the actor ultimately does not obtain.

[16] As Gartrell (1987) remarked, the literature on relative deprivation tends to ignore that ego-centered social networks are a powerful source which determine "who compares

Formally speaking, this second definition of persons "in the same boat," to borrow Stouffer's expression, raises a problem we do not encounter with the "global comparisons" implied by [5] and [6]: how are we to handle the situation where RD^1_{freq} or RD^2_{freq} are nil in the agent's neighborhood? To remain consistent with the posited mechanism, an agent's feeling of envy can only be maximal here (since in this situation he would be the only one who did not get what he wanted). To represent this idea – i.e. "zero neighbors experiencing RD" – I have changed 0 into 0.01 when the situation presents itself (otherwise computation would be impossible), thereby providing maxima that vary with size of agent's neighborhood.[17]

Simultaneously generating RD frequency and RD intensity patterns

The sensitivity analysis that follows aims to demonstrate that within the artificial society driven by the mechanisms just described, the four-way relation between the attractiveness of goods at stake, the "wealth" of the opportunity structure, the quantity of dissatisfied agents (RD^1_{freq} and RD^2_{freq}) and the intensity of this dissatisfaction ($RD^1_{intensity}$ and $RD^2_{intensity}$) assumes multiple forms that are not independent of the interaction configuration linking agents to each other.[18]

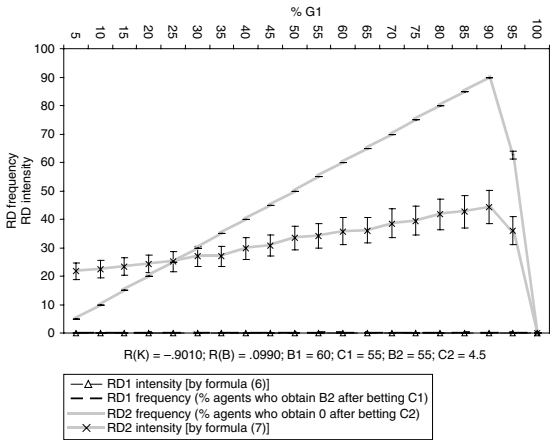
To prove it, I first consider a microcosm without dyadic interactions between agents, and then I introduce these interactions, first in the form of a random network, then a scale-free network.[19]

with whom." In particular, Gartrell (2001: 173–5) demonstrated that dyadic properties such as frequency, "multiplexity" and strength of contacts are especially important in predicting the reference point of a given agent. As I said in my introduction, among formal analyses of relative deprivation, only Burt (1982) explicitly takes into account the role of social networks. In particular, he posits that actors compare with one another if they are structurally equivalent. Assuming that actor's significant others are his direct contacts, I am positing a more general dyadic rule of comparison.

[17] As I noted above, Elster (1999: 141) presents "envy" as a comparison-based emotion, and so did I in Equations [5] and [6]. In Equations [7] and [8], by contrast, where I posit agents to be embedded in a network of dyadic links, "envy" is considered as an interaction-based emotion (emotion that arises "only when there is social interaction," see Elster (1999: 141)).

[18] The attractiveness of G^1 over G^2 is measured by $R(B) = (B^1 - C^1) / (B^2 - C^2)$ and $R(K) = [(B^1 - C^1) - (B^2 - C^2)] / (C^1 - C^2)$ (see respectively, Boudon (1982: 118) and Kosaka (1986: 38)); on the other hand, the "wealth" of the opportunity structure is represented by the percentage of goods with the highest benefit, i.e. goods G^1.

[19] In a previous analysis (see Manzo 2009b), I explored in depth only the relation between the first three elements. The main computational results obtained by analyzing approximately 26,000 parameter combinations concerning both the zero and non-zero-second alternative cases (i.e. respectively, the situation in which $B^2 = C^2 = 0$ and $B^2 \neq 0$ and $C^2 \geq 0$) can be synthesized as follows: (a) the relation between an



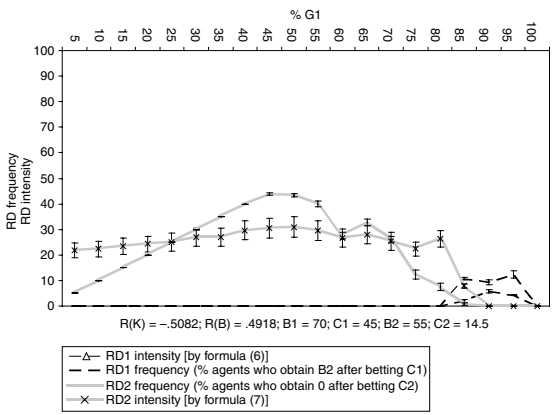
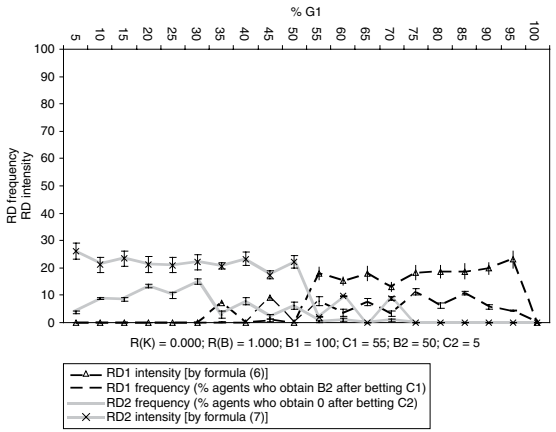
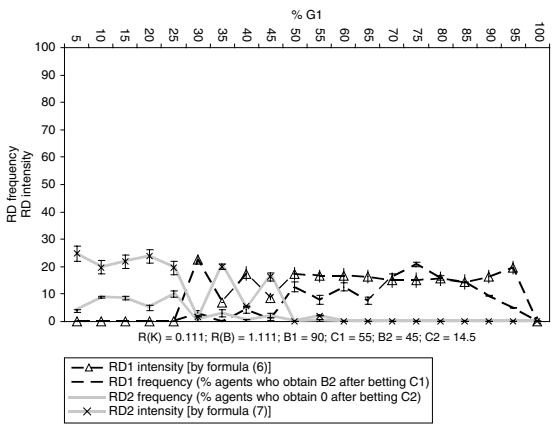


Figure 13.1 Relative deprivation in an artificial society, situation 1 percentages (95% confidence intervals) of agents who finally obtain B2 or nothing after betting C^1 or C^2 (y-axis) and average values of RD^1 and RD^2 intensity (average 95% confidence intervals) for these agents (y-axis) as a function of the percentage of available G^1 (x-axis) and different levels of G^1 attractiveness (see R(K) and R(B) values).





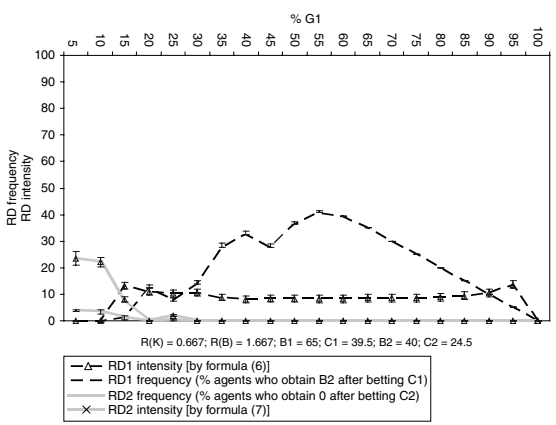
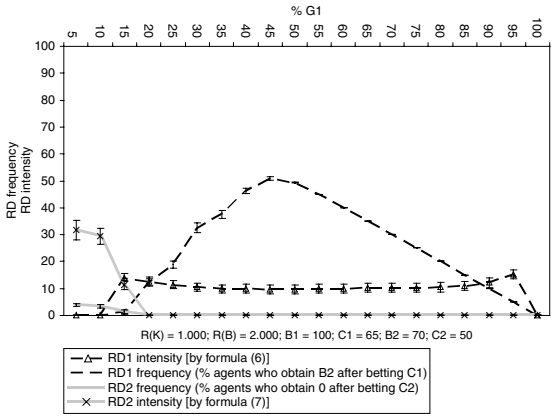
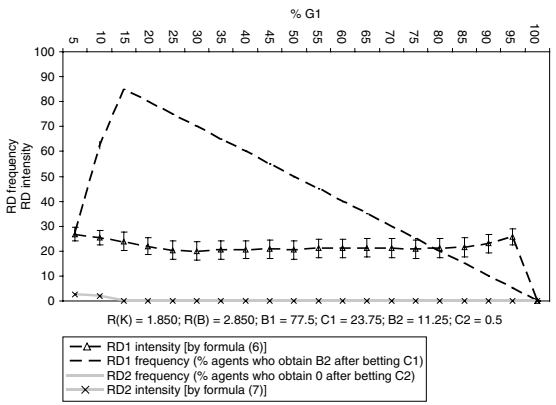


Figure 13.1 (*cont.*)





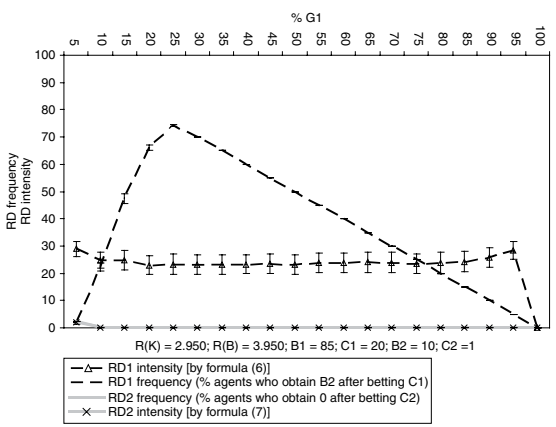
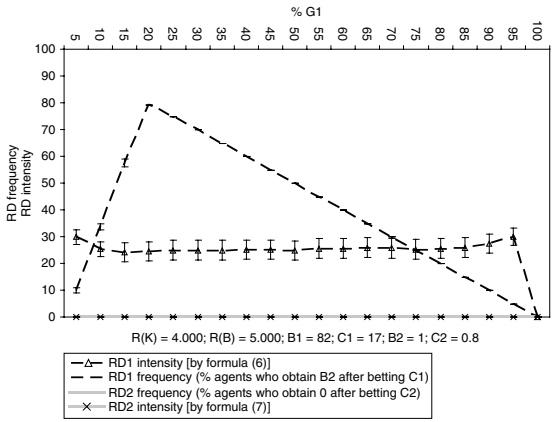
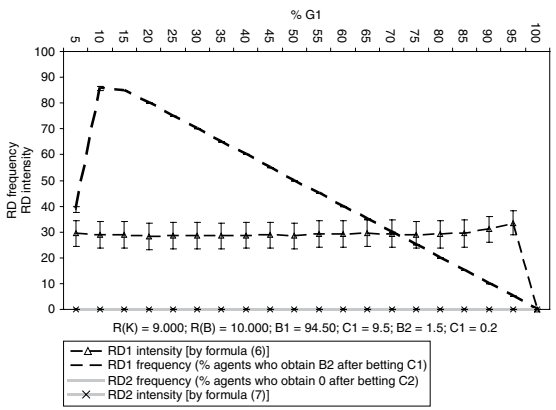


Figure 13.1 (*cont.*)





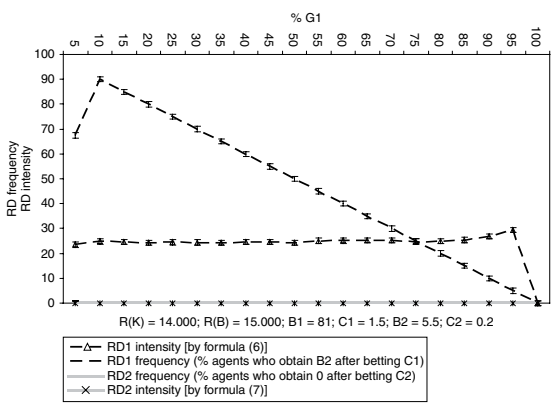


Figure 13.1 (*cont.*)

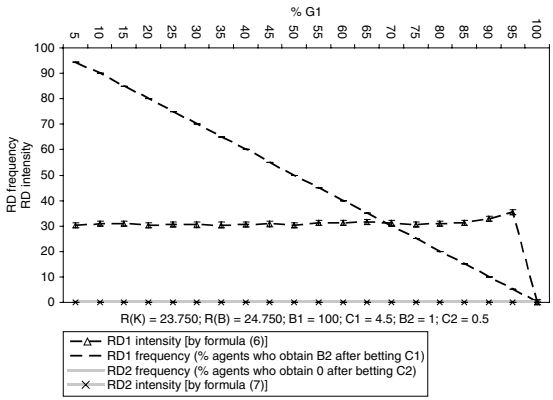


Figure 13.1 (*cont.*)

The population-based inter-individual comparisons case

Figure 13.1 presents the $RD^2_{intensity}$ and $RD^1_{intensity}$ generated by the model for each level of RD^2_{freq} and RD^1_{freq}. To improve graph readability, I have omitted trends in percentages of agents obtaining what they want.[20]

improvement in the opportunity structure and the percentage of dissatisfied agents may take a positive linear form (*more opportunities, more dissatisfied agents*), a negative one (*more opportunities, fewer dissatisfied agents*), or both forms at once; (b) we get closer to the negative linear form (*more opportunities, fewer dissatisfied agents*) as attractiveness of the higher-returning goods makes each agent's choice insensitive to that of others; (c) this multiplicity of forms and the underlying dynamic do not vary relative to population size as long as size change is proportionate to number of higher-returning goods; in the opposite case, the "*more opportunities, more dissatisfied agents*" relation reappears at a rate proportionate to how limited opportunity structure range is relative to population size. These results thus confirm and extend Boudon's, Kosaka's and Yamaguchi's original result (they only studied indeed the situation in which $B^2 = C^2 = 0$), which is that the positive linear form designated by "*more opportunities, more dissatisfied actors*" – i.e. the take-off point of sociological literature on RD – is only validated in a specific region of the parameter space.

[20] I am presenting here a set of typical patterns generated by the model for a specific series of R(B) and R(K) values. In fact, I explored about 1,976 different combinations of B^1, C^1, B^2 and C^2 values (varying respectively between 10 and 100; 1.5 and 95; 5 and 90 and 0 and 50) producing G^1 attractiveness levels ranging from (R(K)) = -0.90 to (R(K)) = 98.5, or, alternatively, (R(B)) = 0.09 to (R(B)) 99.5. Taking into account also the variations of the percentage of G^1 lots, I simulated the model for 20,700 parameter combinations, for a total of 207,000 simulations, since each combination was simulated ten times to assess the model behavior variability linked to its random elements (for the sake of brevity, I omitted here the values of ten seeds I used). All the simulations consider populations of 100 agents demanding a minimal gain of r = 1. This sensitivity analysis was performed using the NetLogo 4.0.3 "BehaviorSpace" module.

We see that RD^2_{freq} and $RD^2_{intensity}$ are more likely to move in the same direction than are the curves relative to RD^1_{freq} and $RD^1_{intensity}$. In the first case, an increase in the proportion of agents aiming for G^2 but getting 0 tends to go together with a more intense feeling of deprivation, and vice versa. In the second case, on the contrary, an increase in the proportion of agents aiming for G^1 but only getting G^2 tends to go together with a less intense feeling of deprivation, whereas that intensity increases when the number of these agents falls.

The aggregate RD^2_{freq} quantity and the individual $RD^2_{intensity}$ thus seem linked by a positive relation ("more more-intensely-dissatisfied individuals" or "fewer less-intensely-dissatisfied individuals"), whereas RD^1_{freq} and $RD^1_{intensity}$ seemed linked by a negative one ("more less-intensely-dissatisfied individuals" or "fewer more-intensely-dissatisfied individuals"). This reflects the fact that the mechanisms I posited as generating the dissatisfaction associated by agents respectively with RD^1 and RD^2 are not the same (compare Equations [5] and [6]).

The case of RD^1 is simple. The feeling of deprivation is assumed here to derive from two sources: a feeling of disappointment whose intensity is proportionate to the size of the gap between expected gain and gain actually realized, and a feeling of envy of an intensity inversely proportional to the rate of deprivation in the population. Under this condition and for a given value of R(K), whereas the value of the term quantifying the first source is stable, the value of the term quantifying the second falls as RD^1_{freq} increases, and vice versa. This means we are first adding a gradually falling quantity, then a gradually rising quantity, to a fixed quantity. In all situations where RD^1_{freq} increases at first and then declines, the result will be a flattened U-curve for $RD^1_{intensity}$. However, as we near the negative form of the relation between opportunity structure and RD^1_{freq}, $RD^1_{intensity}$ will increase more or less slowly, because in this case RD^1_{freq} is only falling.

The case of RD^2 is slightly more complex. Here the feeling of deprivation is understood to derive from a third mechanism, in addition to the other two sources allowed for RD^1; namely, that agents experiencing RD^2 reason counterfactually, and this in turn generates a feeling of regret whose intensity is proportionate to the number of wasted G^1. Under this condition, even though the value of the term quantifying the effect of interpersonal comparisons falls as RD^2_{freq} increases, the value of the term quantifying the effect of the counterfactual reasoning tends to increase. This means that the more abrupt the rise in RD^2_{freq} and the greater its breadth, the more likely it is for a concomitant increase in $RD^2_{intensity}$ to set in. On the other hand, when RD^2_{freq} is low and rises little, $RD^2_{intensity}$ is more likely to be stable (the effects of the two mechanisms cancel each other out) or even to vary inversely with RD^2_{freq} (and

here we come back to the situation characterizing $RD^1_{intensity}$, where the inter-individual comparisons takes precedence).[21]

As soon as we combine the plural forms of the two groups of relations studied thus far – on the one hand, the relation between an improved objective opportunity structure and percentage of dissatisfied agents (RD^2_{freq} and RD^1_{freq}) (see above note 19); on the other, the relation linking percentage of dissatisfied agents with intensity of agents' feelings of deprivation ($RD^2_{intensity}$ and $RD^1_{intensity}$) – the following general result appears: enriching the opportunity structure can indeed maintain a virtuous relation between that structure and both the quota of dissatisfied agents ("more opportunities, fewer dissatisfied individuals") and the intensity of individual feelings of dissatisfaction ("more opportunities, weaker dissatisfaction"). The problem is that if interpersonal comparisons are operative which inversely link the feeling of dissatisfaction to "scarcity" of deprivation experiences, then the regions where these two relations obtain may well fail to overlap. Under these conditions, dissatisfaction intensity will indeed tend to go down when the number of dissatisfied agents increases, whereas when that number falls, dissatisfaction intensity will tend to rise.

The neighborhood-based inter-individual comparisons case

If we now introduce a dyadic tie structure linking agents to each other within the artificial microcosm (see Equations [7] and [8]), how will this change the complex relation between the opportunity structure (here number of G^1), percentage of dissatisfied agents (RD^2_{freq} and RD^1_{freq}), and feeling-of-deprivation intensity ($RD_{intensity}$)?

The relation between G^1 and RD^2_{freq}–RD^1_{freq} should not be affected. In the present version of the model, dyadic agent interactions determine only the set of agents with whom an agent experiencing RD compares himself. $RD^2_{intensity}$ and $RD^1_{intensity}$, therefore, are what may be affected by such interactions.

To see how fully this is confirmed, I put agents into a random network with a slight spatial bias (the average network degree here is 10).

[21] The profile of the curves just discussed and interpreted is stable when we simulate the model after eliminating the source of inter-individual variability I applied to each of the three mechanisms responsible for $RD^1_{intensity}$ and $RD^2_{intensity}$. And their form is not even linked to the range of the three terms that formalize the action of these mechanisms. I also simulated the probabilistic and determinist versions of the model, standardizing each of these terms, by relating it to the difference between its minimal and maximal theoretical values. While this standardized version is unquestionably more elegant formally, it does not change the profile of the curves presented in Figure 13.1, except to further flatten the shape.

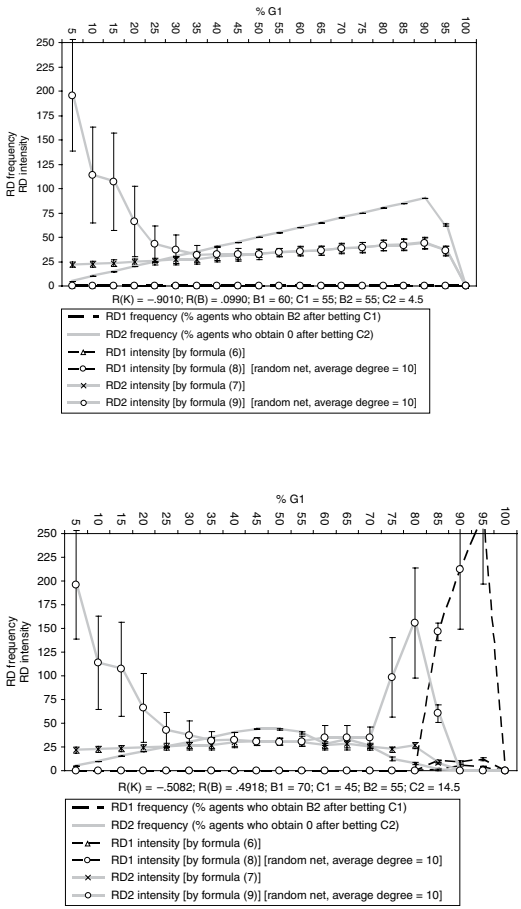


Figure 13.2 Relative deprivation in an artificial society, situation 2 percentages (95% confidence intervals) of agents who finally obtain B2 or nothing after betting C^1 or C^2 (y-axis) and average values of RD^1 and RD^2 intensity (average 95% confidence intervals) for these agents (y-axis) in a no-network world and in a random-network world (average degree = 10) as a function of the percentage of available G^1 (x-axis) and G^1 attractiveness (see R(K) and R(B) values).

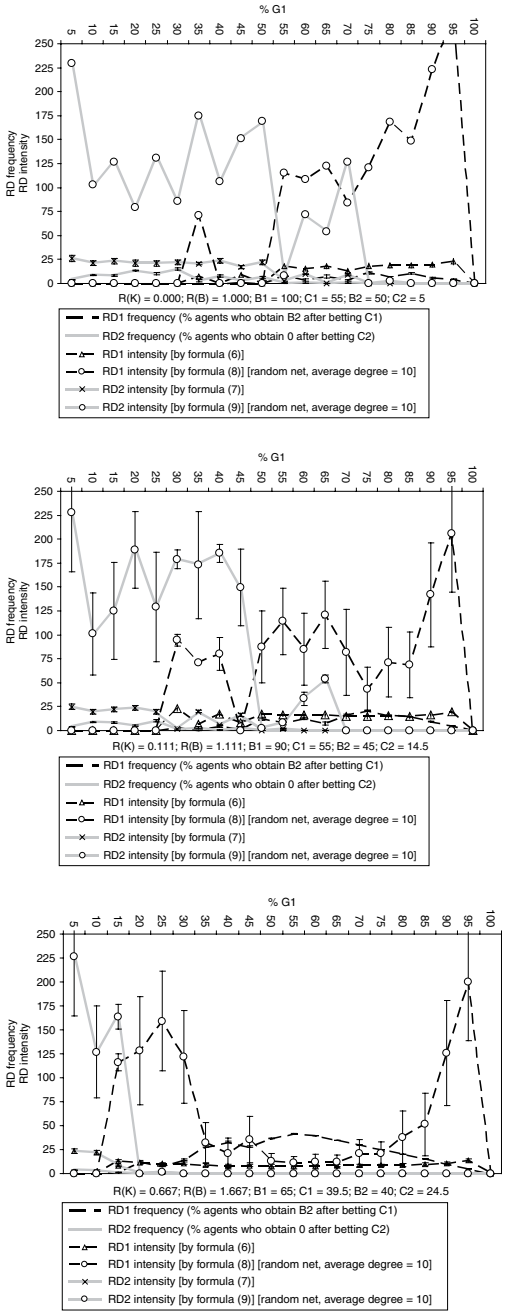
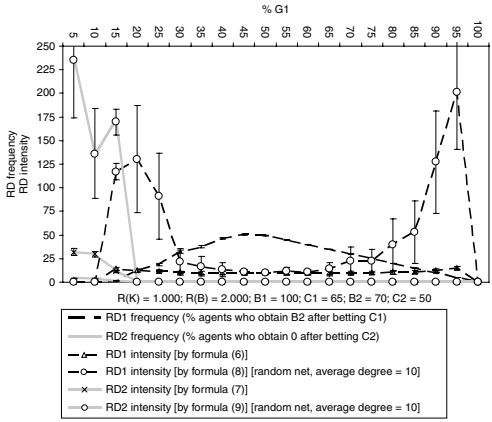
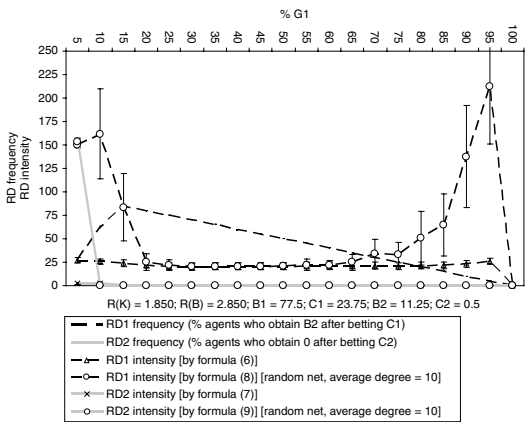


Figure 13.2 (*cont.*)





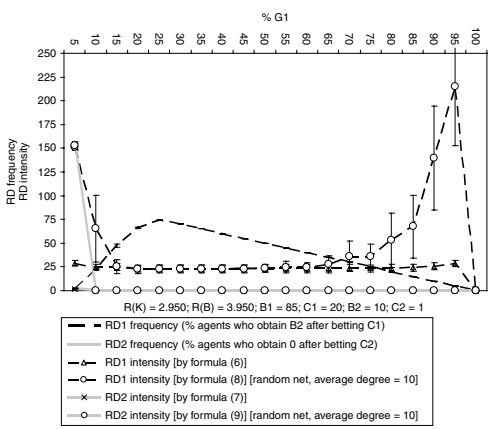


Figure 13.2 (*cont.*)

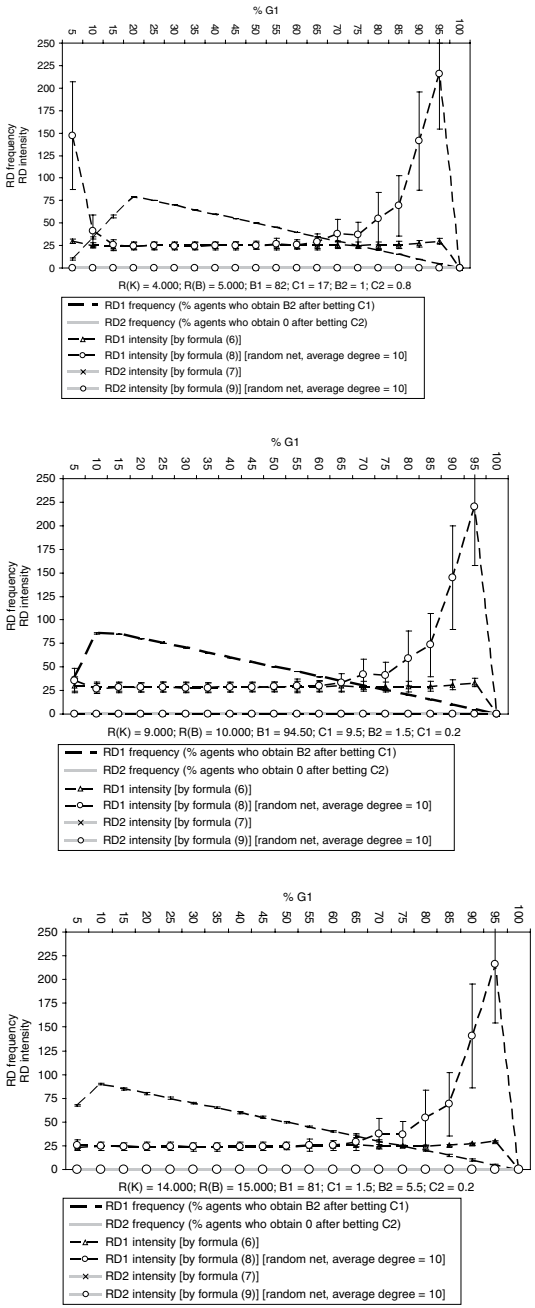


Figure 13.2 (*cont.*)

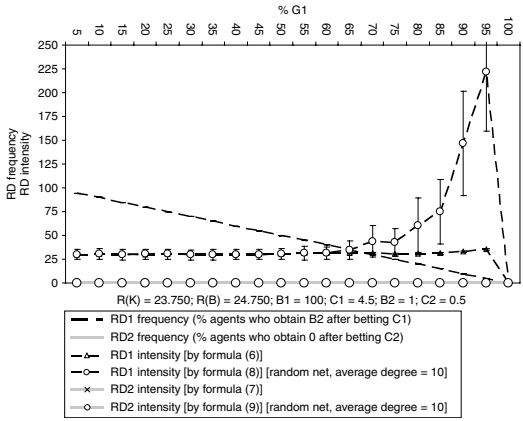


Figure 13.2 (*cont.*)

Figure 13.2 introduces into Figure 13.1 the $RD^2_{intensity}$ and $RD^1_{intensity}$ values I got by simulating the model under this new condition.[22]

We can first consider situations where G^1 attractiveness is weak, leading, as we now know, to an extremely high level of RD^2_{freq}. In this case we observe that in comparison to an artificial world in which there is no network:

1. the relation between an increase in G^1 and $RD^2_{intensity}$ is not exclusively positive but rather both negative and positive;
2. when RD^2_{freq} is low, $RD^2_{intensity}$ level is a great deal higher, but as RD^2_{freq} increases $RD^2_{intensity}$ gets considerably closer to the level observed when there were no ties between agents.

What accounts for these differences? Given that agents embedded in dyadic interactions compare themselves to immediate neighbors experiencing RD^2, the probability of an agent finding other agents around him in the same situation is low when RD^2_{freq} is low. This implies that the term of Equation [7] quantifying the interpersonal comparison will take on extremely high values for many agents, generating a particularly high level of $RD^2_{intensity}$. As RD^2_{freq} increases, this condition disappears: the term relative to interpersonal comparisons will take on increasingly low values while the term relative to counterfactual reasoning will increase continuously. This is why average $RD^2_{intensity}$ levels can fall first, then rise.

[22] To construct the network, I used an algorithm that has only been available in NetLogo since version 4.0.3 (Stonedahl and Wilensky 2008). The algorithm works as follows: (a) we take a randomly chosen agent; (b) we determine the agent closest to him (Euclidian distance); (c) we create a link; (d) we reiterate these operations until the average network degree reaches the average degree chosen at the outset.

If we now consider situations where RD^1_{freq} replaces RD^2_{freq} because G^1 attractiveness is stronger, we observe equally significant modifications. Compared to the no-network artificial world, the form of the relation between an increase of G^1 and $RD^1_{intensity}$ does not change – we move gradually from a mixed negative/positive relation to an entirely positive one ("more opportunities, stronger dissatisfaction") – but the levels of $RD^1_{intensity}$ are much higher at the extremes; that is, when RD^1_{freq} is low. This is because, here again, the overall "scarcity" of RD^1 implies the presence of many "neighborhoods" in which agents experiencing RD^1 have no neighbors in this same deprived situation. Since this agent is the only one not to get what he wanted, he feels maximum envy.[23]

As Figure 13.3 shows, that this structural condition exists is attested by the results of simulations where the average network degree went from 10 to 50. Under this condition, the differences in average $RD^1_{intensity}$ and $RD^2_{intensity}$ levels that existed between societies with and societies without random networks tend to disappear. This is because given that agents' "neighborhoods" have been extended, an agent in RD^1 or RD^2 is more likely to meet someone among the neighbors he is linked to who is also experiencing RD, despite the fact that the overall rate of RD^1 and RD^2 are low. The effect is to contain quite firmly the spectacular rise of the term quantifying neighborhood-based inter-individual comparisons.

We can obtain more direct proof of this phenomenon by introducing a scale-free network (rather than a random one) into the model. The purpose of doing this is to construct by default a situation with a great number of small "neighborhoods," thereby structurally multiplying situations where an agent experiencing RD^1 or RD^2 is unlikely to find another in the same situation. This should greatly amplify average $RD_{intensity}$ levels.[24]

Figure 13.4 show that this is exactly what happens. Regardless of G^1 attractiveness, $RD^2_{intensity}$ or $RD^1_{intensity}$ are indeed regularly higher in the artificial society based on a scale-free network than they are in

[23] This aggregate effect will, of course, appear more or less sharp depending on the functional form chosen to formalize the term quantifying interpersonal comparisons in the case where agent has no neighbors in the same deprivation situation as his. It can be almost entirely effaced, for example, by applying a logarithmic transformation to this term in Equations [7] and [8]. Having studied the behavior of the model under that alternative condition, I conclude, however, that this type of manipulation conceals the presence of a significant theoretical phenomenon.

[24] To construct this network I used an algorithm available in NetLogo (Wilensky 2005) based on a formalization of the "preferential attachment" mechanism first put forward by Barabasi and Réka (1999). Researchers are currently at work constructing algorithms formalizing mechanisms that will generate scale-free (and small-world)networks that are sociologically more significant than the one implemented with the algorithm I used (see, for instance, Pujol *et al.* 2005). Here, however, what interests me are the structural characteristics of a scale-free network, not the process by which it emerges.

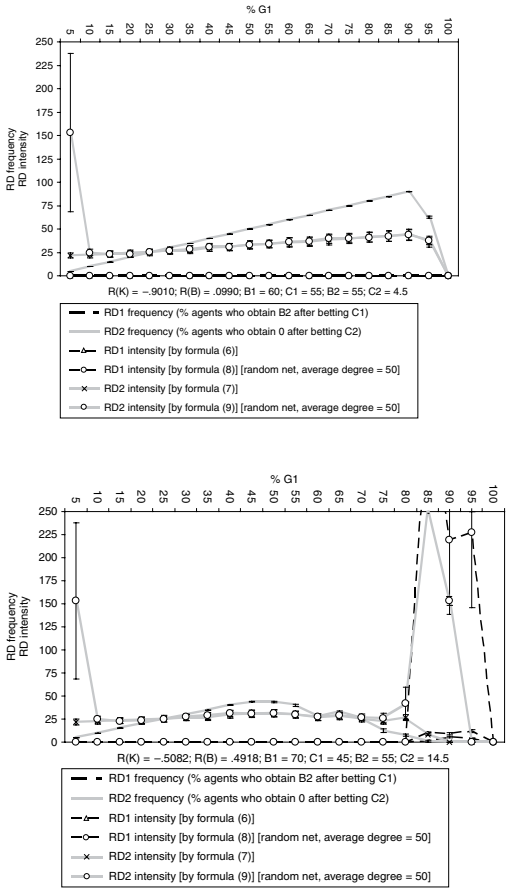


Figure 13.3 Relative deprivation in an artificial society, situation 3 percentages (95% confidence intervals) of agents who finally obtain B^2 or nothing after betting C^1 or C^2 (y-axis) and average values of RD^1 and RD^2 intensity (average 95% confidence intervals) for these agents (y-axis) in a no-network world and in a random-network world (average degree = 50) as a function of the percentage of available G^1 (x-axis) and G^1 attractiveness (see R(K) and R(B) values).

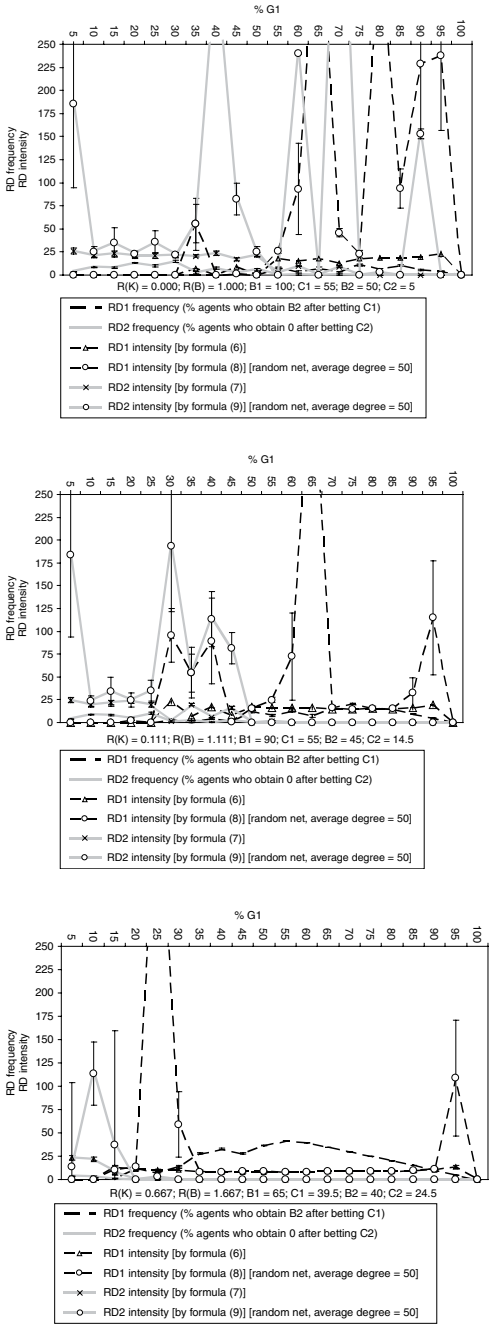
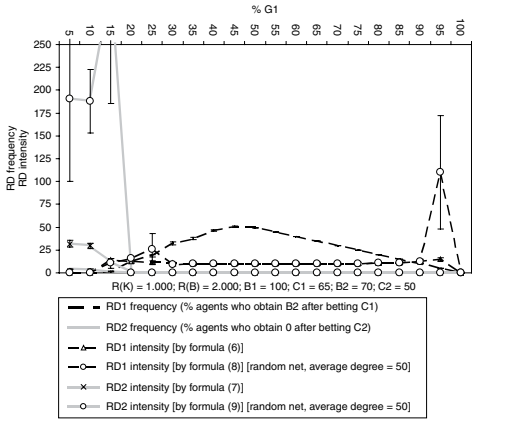
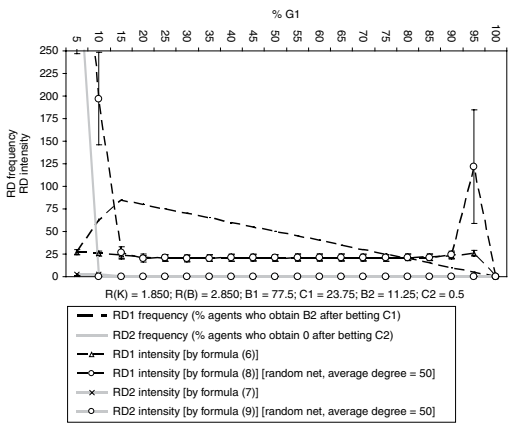


Figure 13.3 (*cont.*)





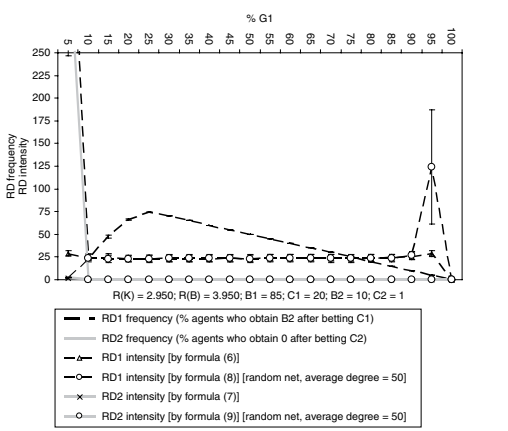


Figure 13.3 (*cont.*)

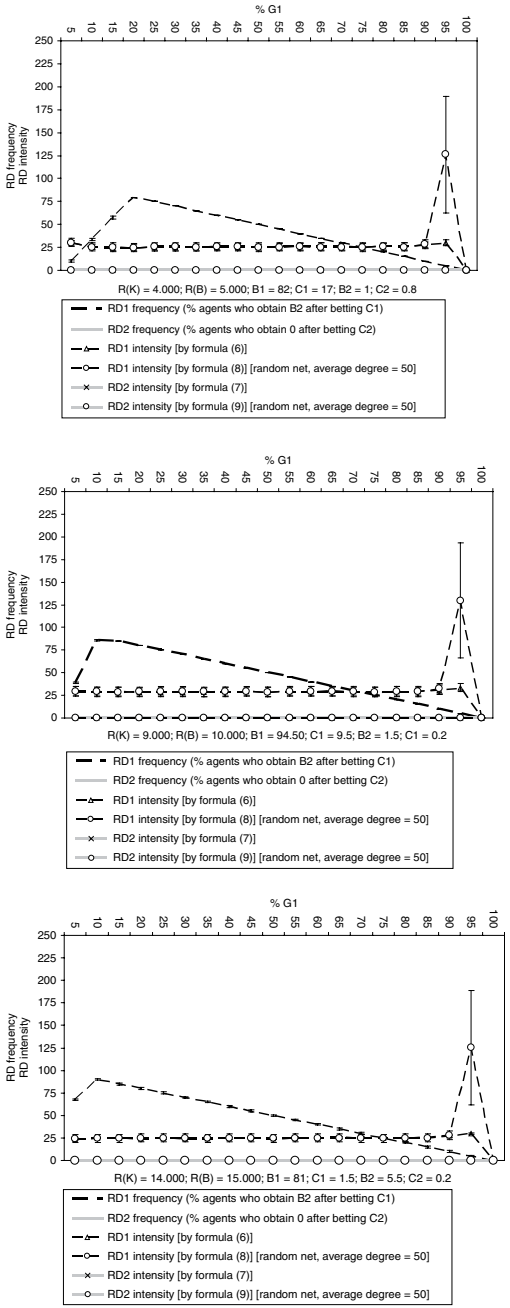


Figure 13.3 (*cont.*)

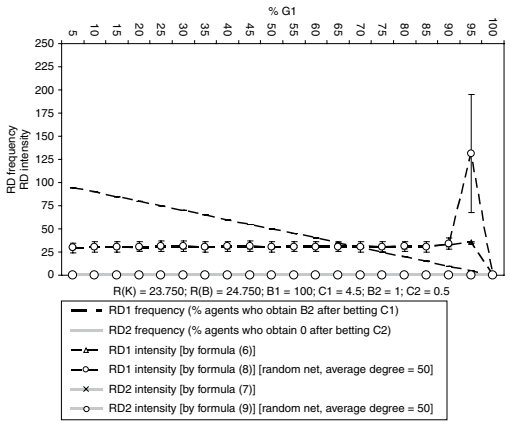


Figure 13.3 (*cont.*)

an artificial society without dyadic interactions that delimit the set of agents with whom one compares oneself. The same is true if we compare the scale-free network microcosm with the random network society (see Figure 13.2), except for the extreme situations, i.e. where RD^2_{freq} or RD^1_{freq} is low.

This is readily explained. Though there are indeed a great many small "neighborhoods" with the scale-free network – and this increases the probability that agents experiencing RD will not meet another agent in their neighborhood who is also experiencing RD – these same limited neighborhoods mean that the agent is alone among a low number of satisfied fellow agents. Despite the fact that dissatisfaction is maximal here compared to the situation where one has at least a few neighbors who are also experiencing RD, this maximum will be lower compared to the random-network artificial society (comprising an average degree of 10) where one may be alone among a higher number of satisfied agents.

Tables 13.1 and 13.2 directly demonstrate (for the two extreme values of R(B)–R(K)) these structural bases of the differences in average $RD^2_{intensity}$ or $RD^1_{intensity}$ levels that emerge in the artificial society with a random network and the society with a scale-free network. In the simulations just commented on, we see on the one hand that degree of agents experiencing RD^2 and RD^1 is on average lower in the scale-free network than in the near-random one, and on the other, the percentage of agents who are the only ones in their neighborhood experiencing RD is on average higher in the first case than in the second.

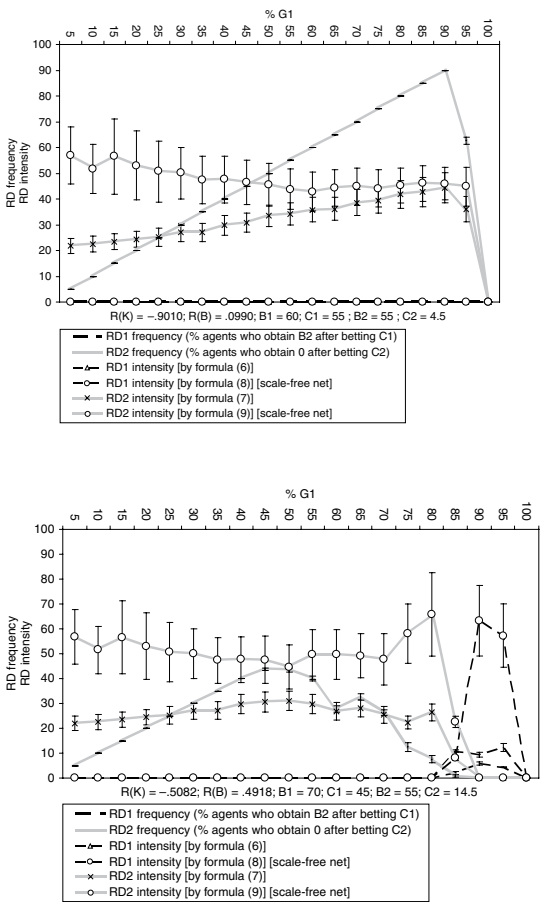


Figure 13.4 Relative deprivation in an artificial society, situation 4 percentages (95% confidence intervals) of agents who finally obtain B2 or nothing after betting C^1 or C^2 (y-axis) and average values of RD^1 and RD^2 intensity (average 95% confidence intervals) for these agents (y-axis) in a no-network world and in a scale-free-network world as a function of the percentage of available G^1 (x-axis) and G^1 attractiveness (see R(K) and R(B) values).

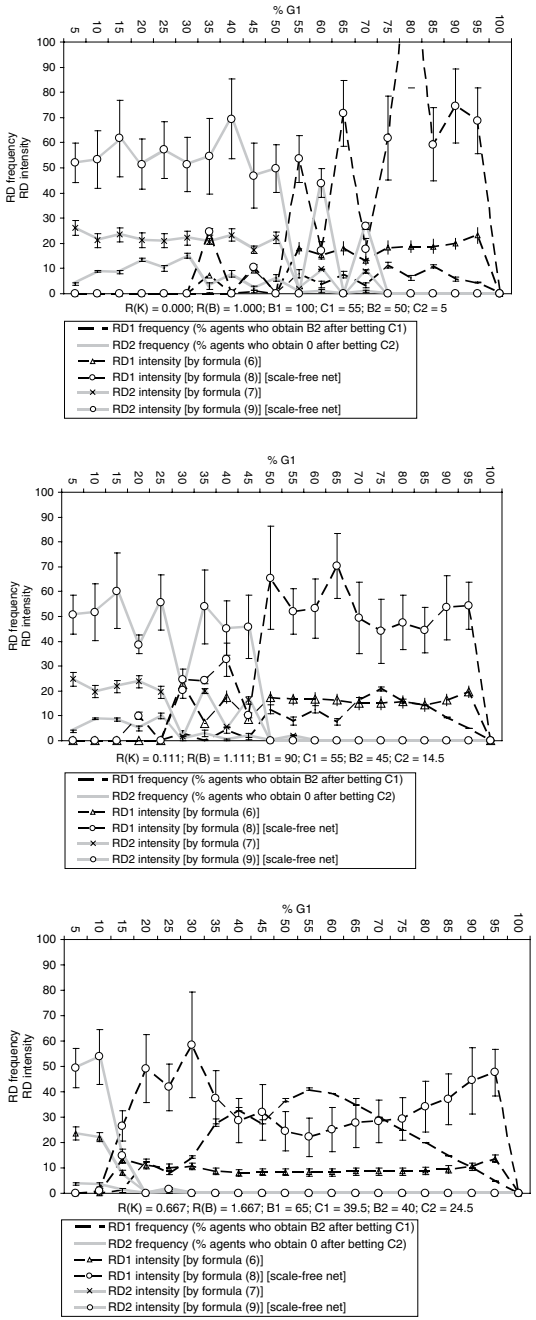
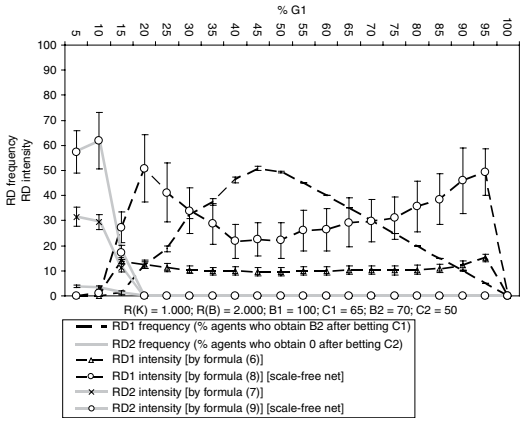
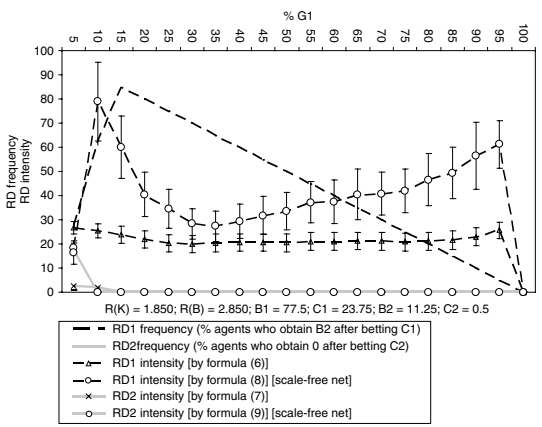


Figure 13.4 (*cont.*)





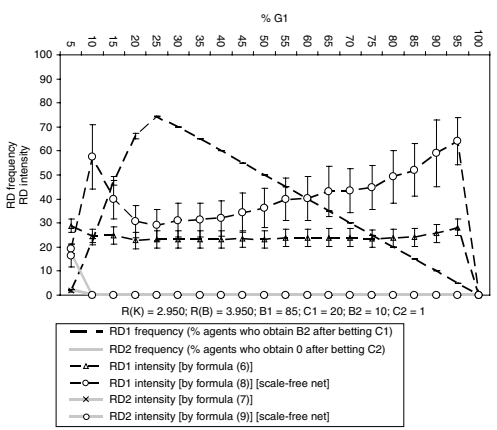


Figure 13.4 (*cont.*)

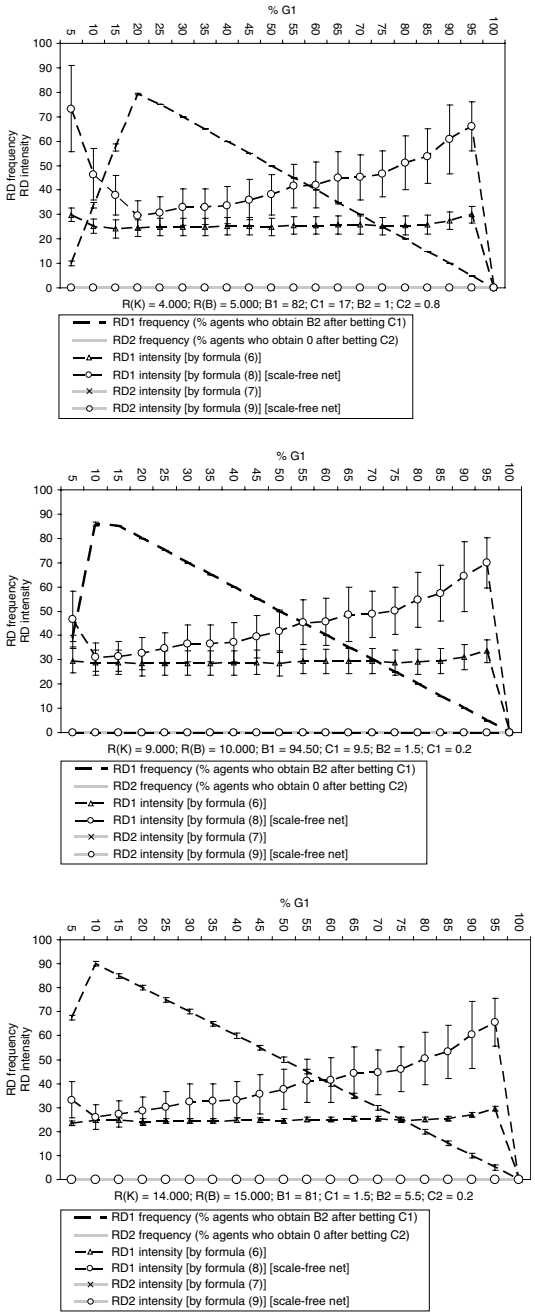


Figure 13.4 (*cont.*)

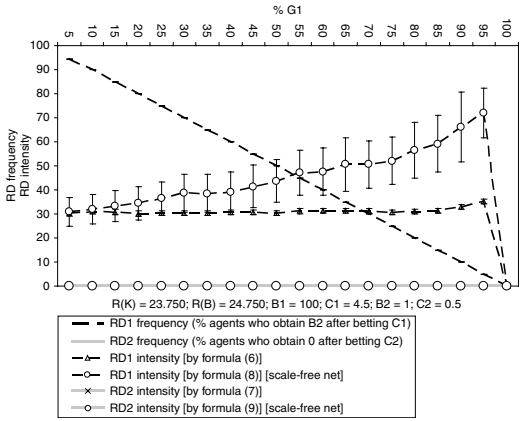


Figure 13.4 (*cont.*)

These results suggest that dyadic interactions matter. In my minimalist hypothetical schema – my only assumption was that network influences actors' points of comparison – these simulations indicate that the presence of interactions can significantly modify the individual dissatisfaction levels which characterize a microcosm whose agents are entirely isolated. Restricting the bases for inter-individual comparisons amounts, paradoxically, to increasing the probability of individual dissatisfaction being stronger. The more the dyadic interaction configuration multiplies the number of neighborhoods where the agent is the only one not to have what he wants, the further we move away from the dissatisfaction levels that appear for an artificial world where the agent compares his deprivation situation to the global diffusion of deprivation in the population at large.

Concluding remarks

This chapter has aimed to sketch a unified theoretical framework which links two classes of problems: on the one hand, the one of simultaneously generating variable quantities of dissatisfied actors and heterogeneously intense individual feelings of dissatisfaction; on the other hand, the one of determining how this dissatisfaction is modified when, instead of taking into account overall success rates, individuals only consider the success rate of their closest contacts.

Methodologically, the point here has been to suggest that this undertaking can now benefit from a computational tool – agent-based

Table 13.1 *Average degree of agents experiencing RD² and percentage of these agents who do not have any neighbors in RD² (average values with standard deviation in parentheses) for each of the three network structures used (case of R(K) = -0.09. See Figures 13.2, 13.3 and 13.4 for RD frequency and RD intensity trends)*

	Random network (average degree = 10)		Random network (average degree = 50)		Scale-free network	
G1	RD² agents' average degree	% of agents who do not have any neighbors in RD²	RD² agents' average degree	% of agents who do not have any neighbors in RD²	RD² agents' average degree	% of agents who do not have any neighbors in RD²
5	9.9 (1.5)	52.0 (24.0)	52.5 (5.2)	0.0	1.7 (0.8)	88.0 (18.3)
10	10.2 (0.8)	35.0 (18.6)	51.2 (3.3)	0.0	1.7 (0.7)	80.0 (15.5)
15	9.9 (0.7)	16.7 (10.4)	50.9 (3.3)	0.0	2.1 (1.0)	74.0 (18.2)
20	9.9 (0.5)	13.0 (6.0)	49.7 (2.6)	0.0	2.0 (0.8)	71.5 (14.8)
25	9.9 (0.5)	10.0 (6.5)	49.9 (2.2)	0.0	1.9 (0.6)	66.0 (13.4)
30	9.9 (0.4)	4.7 (3.4)	49.6 (2.6)	0.0	2.0 (0.5)	59.0 (12.3)
35	10.0 (0.24)	1.7 (1.9)	49.0 (2.3)	0.0	2.0 (0.4)	52.0 (9.7)
40	10.0 (0.2)	0.5 (1.0)	49.5 (2.1)	0.0	2.0 (0.3)	49.8 (12.2)
45	10.0 (0.2)	0.9 (1.5)	49.9 (2.0)	0.0	2.0 (0.2)	42.4 (8.6)
50	10.0 (0.2)	0.0	49.7 (1.4)	0.0	2.0 (0.2)	36.4 (7.3)
55	10.0 (0.2)	0.0	50.0 (1.1)	0.0	2.1 (0.2)	28.4 (7.1)
60	10.0 (0.2)	0.0	49.9 (1.2)	0.0	2.1 (0.1)	25.8 (7.7)
65	10.0 (0.2)	0.0	49.9 (1.0)	0.0	2.1 (0.1)	22.3 (4.8)
70	10.0 (0.1)	0.0	49.8 (0.8)	0.0	2.1 (0.1)	20.6 (4.5)
75	10.0 (0.1)	0.0	49.9 (0.7)	0.0	2.1 (0.1)	17.7 (4.1)
80	10.0 (0.1)	0.0	49.8 (0.8)	0.0	2.1 (0.1)	13.8 (5.5)
85	10.0 (0.1)	0.0	49.7 (0.6)	0.0	2.1 (0.1)	9.4 (3.8)
90	10.0 (0.1)	0.0	49.8 (0.5)	0.0	2.1 (0.1)	6.8 (4.1)
95	9.9 (0.2)	0.0	49.7 (0.8)	6.0 (12.8)	2.1 (0.1)	25.3 (7.9)
100	—	—	—	—	—	—

Table 13.2 *Average degree of agents experiencing RD[1] and percentage of these agents who do not have any neighbors in RD[1] (average values with standard deviation in parentheses) for each of the three network structures used (case of R(K) = 23.75. See Figures 13.2, 13.3 and 13.4 for RD frequency and RD intensity trends)*

	Random network (average degree = 10)		Random network (average degree = 50)		Scale-free network	
G1	RD[1] agents' average degree	% of agents who do not have any neighbors in RD[1]	RD[1] agents' average degree	% of agents who do not have any neighbors in RD[1]	RD[1] agents' average degree	% of agents who do not have any neighbors in RD[1]
5	10.0 (0.1)	0.0	50.0 (0.3)	0.0	2.0 (0.1)	3.7 (2.4)
10	10.0 (0.1)	0.0	50.0 (0.3)	0.0	2.0 (0.1)	7.2 (4.3)
15	10.0 (0.1)	0.0	49.7 (0.5)	0.0	2.0 (0.1)	9.6 (5.6)
20	10.0 (0.1)	0.0	49.7 (0.7)	0.0	2.0 (0.1)	12.6 (6.5)
25	10.1 (0.2)	0.0	49.7 (0.7)	0.0	2.0 (0.2)	18.5 (9.4)
30	10.1 (0.2)	0.0	50.2 (0.9)	0.0	2.0 (0.2)	22.4 (10.4)
35	10.1 (0.2)	0.0	50.3 (1.1)	0.0	2.0 (0.2)	24.3 (10.8)
40	10.1 (0.2)	0.0	50.2 (1.2)	0.0	2.0 (0.2)	26.0 (10.5)
45	10.1 (0.3)	0.0	50.1 (1.2)	0.0	2.0 (0.3)	32.0 (14.5)
50	10.1 (0.2)	0.4 (0.8)	50.1 (0.9)	0.0	2.0 (0.4)	37.6 (17.1)
55	10.3 (0.2)	0.4 (0.9)	50.0 (1.7)	0.0	2.0 (0.4)	41.6 (16.0)
60	10.3 (0.3)	0.8 (0.1)	49.8 (2.0)	0.0	2.1 (0.4)	44.0 (12.9)
65	10.4 (0.3)	1.4 (1.8)	49.7 (2.0)	0.0	2.0 (0.5)	50.0 (14.2)
70	10.4 (0.4)	1.7 (1.7)	49.6 (2.5)	0.0	2.1 (0.6)	55.3 (16.6)
75	10.3 (0.4)	5.6 (3.7)	49.5 (3.2)	0.0	2.1 (0.7)	57.6 (13.5)
80	10.4 (0.4)	9.0 (6.2)	49.6 (3.2)	0.0	1.8 (0.6)	70.5 (14.2)
85	10.5 (0.5)	18.7 (8.8)	49.6 (3.3)	0.0	1.8 (0.5)	72.7 (16.5)
90	10.4 (0.7)	32.0 (19.4)	49.1 (4.6)	0.0	1.8 (0.6)	78.0 (18.3)
95	10.3 (1.1)	60.0 (32.2)	50.0 (0.3)	10.0 (13.41)	1.5 (0.3)	92.0 (16.0)
100	—	—			—	—

modeling – that allows both for specifying in a highly flexible way the conceptual structure of a bundle of mechanisms and exploring their aggregate causal effects under a large range of conditions.

Introducing at the same time generative mechanisms of relative deprivation rate and feelings thus allowed to establish that an improving opportunities system may go along with two different situations. On the one hand, it can produce a "*more opportunities, more dissatisfied-*

yet-less-intensely-dissatisfied agents" pattern; on the other hand, it may go together with a "*more opportunities, fewer dissatisfied-yet-more-intensely-dissatisfied agents*" pattern. The condition under which the model studied here leads to the emergence of these complex relations is the presence of interpersonal comparisons that inversely tie individual dissatisfaction to the diffusion of deprivation situations.

The theoretical interest of these computational results is that they circumscribe the extension of the classic "*more opportunities, higher dissatisfaction levels*" pattern, showing that the inverse pattern, i.e. "*more opportunities, lower dissatisfaction levels*," is equally possible. However, they also signal that the two patterns might be incompatible. This incompatibility is exemplified in the extreme case where all actors want to obtain the most attractive goods regardless of how many competitors they think they have. In this case, as opportunities improve, the quantity of dissatisfied agents falls while the intensity of deprived agents' dissatisfaction can only grow. According to the absolute intensity of this feeling, society's level of individual dissatisfaction could ultimately fall (if these agents are not intensely dissatisfied) or, on the contrary, rise (if these agents, while few in number, are also intensely dissatisfied).

Dyadic interaction configuration then can play a decisive role in the appearance of one or another systemic equilibrium. The last variant of the model simulated here suggests that if we suppose that actors take account of deprivation diffusion within their local neighborhood rather than throughout the population, individual dissatisfaction levels tend to soar. If the network contains few low-degree-nodes, this explosion is primarily verified when the global quota of dissatisfied agents is reduced; by contrast, the rise tends to become general if there are many low-degree-nodes. In this case, regardless of the global proportion of dissatisfied agents, the probability that each will be the only one among his contacts not to have what he wants rises.

The theoretical interest of introducing several structures of dyadic ties, which has been greatly facilitated by agent-based modeling, is considerable. First, even though the idea is an old one – Merton had already distinguished comparisons of self "with those men who are in some pertinent respect of the same status or in the same category" from comparisons "with the situation of others with whom [one was] in actual association, in sustained social relations" (Merton 1957: 231) – a formal model of relative deprivation implementing this distinction was still missing. This seems to represent real progress because, as (Gartrell 1987: 49) noticed, "the network approach will help to resolve fundamental, unanswered questions about social evaluation first raised in 1950 by Merton and Rossi – specifically, the

origins of comparative frameworks and the relation between individual and categorical or group reference points." Second, introducing neighborhood-based comparisons gives us the occasion to refine some existing conceptual distinctions. On the one hand, insofar as the last version of the model conceives "envy" as a by-product of comparisons that are driven by dyadic links between actors, it seems reasonable to introduce a hybrid category, i.e. what one can call "comparison–interaction-based emotions," in Elster's (1999: 141–2) original typology, which distinguishes between comparison-based emotions and interaction-based emotions (see above note 17). On the other hand, this concept tends to make more complex Hedström's (2005: ch. 3, fig. 3.2) typology of social interactions. In addition to "desire-mediated," "belief-mediated" and "opportunity-mediated interactions," we should indeed also take into account the possibility of "emotion-mediated interactions."

The main limitations of the results discussed are equally obvious. First, the model presented here is still excessively simple compared with the mechanisms which we imagine generate both dissatisfied actor rates in a given society and the intensity of their feelings of dissatisfaction. Second, whatever the degree of theoretical complexity we grant these mechanisms, we would have to demonstrate that they are operative in real societies. What I was looking for in this preliminary analysis was simply material that would serve to convince the reader that agent-based simulation constitutes a particularly well-adapted tool for analytical sociology, enabling sociologists in this field to study as completely as possible the causal implications of the models they aspire to analyze.

With regard to theoretical enrichment, it should be evident that this type of analysis can be pushed as far as we like. But the technique can also be extremely useful when the objective is to link the model to reality. It is perfectly capable of handling fine empirical data on reasoning, comparisons, feelings and/or specific objects grounded in individual states of deprivation. Likewise, the regularities discussed here can easily be compared with empirical quantifications of dissatisfied actor rates as well as of components of the individual feeling of dissatisfaction.

From this perspective, agent-based models offer an additional general benefit: they pinpoint just where our empirical data are deficient, thereby suggesting how to reorient our collecting procedures.

So let us conclude this chapter by emphasizing the importance of pursuing that connection between analytical sociology, agent-based models and more traditional quantitative techniques.

In the minds of many quantitative sociologists, analytical sociology is merely an empty shell. Morgan (2005: 26), for example, wrote as follows of the contributions assembled by Hedström and Swedberg (1998): "Without a doubt, they correctly identified a major problem with quantitatively oriented sociology. But, they did not offer a sufficiently complete remedy." Pisati (2007: 7–8) recently wrote: "It is not clear how the explanation strategy in question can be applied in practice to explain complex systems – which is what social phenomena constantly are."

Though we can agree that "there is no method, let alone a logic, for conjecturing mechanisms … this is an art, not a technique" (Bunge 2004: 200), it seems to me urgent to have the "art" give way to agent-based modeling when we study such mechanisms. If we cannot resolve to take this step, then this false definition of the situation – i.e. "analytic sociology is an empty shell" – could become true.

REFERENCES

Axelrod, Robert. 1997. *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration*. Princeton University Press.

Axtell, Robert. 2000. "Why agents? On the varied motivations for agent computing in the social sciences," in Robert Axtell (ed.), *Proceedings of the Workshop on Agent Simulation: Applications, Models and Tools*, Argonne, IL, Argonne National Laboratory, 3–24.

Barabasi, Albert-Laszlo and Albert Réka. 1999. "Emergence of scaling in random networks," *Science* 286(5439): 509–12.

Boudon, Raymond. 1979. "Generating models as a research strategy," in Robert K. Merton, James S. Coleman and Peter H. Rossi (eds.), *Qualitative and Quantitative Social Research*. New York: The Free Press, 51–64.

1982 [1977]. *The Unintended Consequences of Social Action*. London: Macmillan.

Bruch, Elizabeth and Robert D. Mare. 2006. "Neighborhood choice and neighborhood change," *American Journal of Sociology* 112(3): 667–709.

Bunge, Mario. 2004. "How does it work? The search for explanatory mechanisms," *Philosophy of the Social Sciences* 34(2): 260–82.

Burt, Ronald S. 1982. *Toward a Structural Theory of Action*. New York: Academic Press.

Cederman, Lars-Erik. 2001. "Agent-based modelling in the political science," *Political Methodology* 10(1): 16–22.

2005. "Computational models of social forms: advancing generative process theory," *American Journal of Sociology* 110(4): 864–93.

Cherkaoui, Mohamed. 2001. "Relative deprivation," in *The International Encyclopaedia of Social and Behavioral Sciences*. London: Elsevier, 2002, 3522–6.

2005. *Invisible Codes. Essays on Generative Mechanisms*. Oxford: Bardwell Press.

Clark, Andrew, Paul Frijters and Michael Shields. 2008. "Relative income, happiness and utility: an explanation for the Easterlin paradox and other puzzles," *Journal of Economic Literature* 46(1): 95–144.
Coleman, James S. 1990. *The Foundations of Social Theory*. Cambridge, MA: Harvard University Press.
Cox, David R. 1992. "Causality: some statistical aspects," *Journal of the Royal Statistical Society*, Series A, 155(2): 291–301.
Crosby, F.J. 1976. "A model of egoistical relative deprivation," *Psychological Review* 83: 85–113.
Davis, J.A. 1959. "A formal interpretation of the theory of relative deprivation," *Sociometry* 22: 280–96.
Durkheim, Émile. 1951 [1897]. *Le suicide*. Paris: PUF (here, English trans. John A. Spaulding and George Simpson, New York: Simon and Schuster, 1951).
Easterlin, Richard. 1973. "Does money buy happiness?" *The Public Interest* 3 (Winter): 3–10.
Elster, Jon. 1999. *Alchemies of the Mind: Rationality and the Emotions*. Cambridge University Press.
2007. *Explaining Social Behaviour: More Nuts and Bolts for the Social Sciences*. Cambridge University Press.
Epstein, Joshua. 2006. *Generative Social Science: Studies in Agent-Based Computational Modeling*. Princeton University Press.
Fararo, T.J. 1989. *The Meaning of General Theoretical Sociology. Tradition and Formalization*. Cambridge University Press.
2009. "Generativity," in M. Cherkaoui and P. Hamilton (eds.), *Boudon: a Life in Sociology*. Oxford: Bardwell Press.
Ferber, J. 1999. *Multi-Agent Systems. An Introduction to Distributed Artificial Intelligence*. London: Addison Wesley.
Gambetta, Diego. 1998. "Concatenations of mechanisms," in Peter Hedström and Richard Swedberg (eds.), *Social Mechanisms. An Analytical Approach to Social Theory*. Cambridge University Press, 102–24.
Gartrell, David. 1987. "Network approaches to social evaluation," *Annual Review of Sociology* 13: 49–66.
2001. "The embeddedness of social comparison," in Iain Walker and Heather Jean Smith (eds.), *Relative Deprivation. Specification, Development and Integration*. Cambridge University Press.
Gilbert, G. Nigel. 2007. *Agent-Based Models*. London: Sage Publications.
Goldthorpe, J.H. 2001. "Causation, statistics, and sociology," *European Sociological Review* 17(1): 1–20.
Gurr, Ted Robert. 1970. *Why Men Rebel*. Princeton University Press.
Harré, R. 1972. *The Philosophies of Science. An Introductory Survey*. Oxford University Press.
Hedström, P. 2004. "Generative models and explanatory research: on the sociology of Aage B. Sørensen," *Research in Social Stratification and Mobility* 21: 13–25.
2005. *Dissecting the Social: On the Principles of Analytical Sociology*. Cambridge University Press.
Hedström, P. and P. Bearman. 2009. "What is analytical sociology all about?" in P. Bearman and P. Hedström (eds.), *The Oxford Handbook of Analytical Sociology*. Oxford University Press.

Hedström, P. and R. Swedberg (eds.) 1998. "Social mechanisms: an introductory essay," in P. Hedström and R. Swedberg (eds.), *Social Mechanisms. An Analytical Approach to Social Theory*. Cambridge University Press.

Hedström, P. and L. Udéhn. 2009. "Analytical sociology and theories of middle-range," in P. Bearman and P. Hedström (eds.), *The Oxford Handbook of Analytical Sociology*. Oxford University Press.

Hedström, P. and P. Ylikoski. 2010. "Causal mechanisms in the social sciences," *Annual Review of Sociology* 36: 49–67.

Hummon, Norman P. and Thomas J. Fararo. 1995. "The emergence of computational sociology," *Journal of Mathematical Sociology* 20(2–3): 78–87.

Janssen, Marco, Lilian Naia Aless, Michael Barton, Sean Bergin and Allen Lee. 2008. "Towards a community framework for agent-based modelling," *Journal of Artificial Societies and Social Simulation* 11(2).

Jasso, Guillemina. 2008. "A new unified theory of sociobehavioral forces," *European Sociological Review* 24(4): 411–34.

Kosaka, Kenji. 1986. "A model of relative deprivation," *Journal of Mathematical Sociology* 12(1): 35–48.

Lewis, D. 1973. "Causation," *Journal of Philosophy* 70: 556–67.

Lundquist, J.H. 2008. "Ethnic and gender satisfaction in the military: the effect of a meritocratic institution," *American Sociological Review* 73: 477–96.

Machamer, P.K., L. Darden and C.F. Craver. 2000. "Thinking about mechanisms," *Philosophy of Science* 67(1): 1–25.

Macy, Michael W. and Robert Willer. 2002. "From factors to actors: computational sociology and agent-based modeling," *Annual Review of Sociology* 28: 143–66.

Macy, M.W. and A. Flache. 2009. "Social dynamics from the bottom up: agent-based models of social interaction," in P. Hedström and P. Bearman (eds.), *The Oxford Handbook of Analytical Sociology*. Oxford University Press.

Manzo, Gianluca. 2007a. "Variables, mechanisms and simulations: can the three methods be synthetized?" *Revue française de sociologie* 48, Supplement: 35–71.

2007b. "Comment on Andrew Abbott," *Sociologica* 2.

2009a. *La spirale des inégalités. Choix scolaires en France et en Italie au XX^e^ siècle*. Paris: Presses de l'Université Paris-Sorbonne.

2009b. "Boudon's model of relative deprivation revisited," in M. Cherkaoui and P. Hamilton (eds.), *Raymond Boudon: a Life in Sociology*. Oxford: Bardwell Press, vol. III, part 3, ch. 46, 91–121.

2010. "Analytical sociology and its discontents," *Archives Européennes de Sociology/European Journal of Sociology* 51(1): 129–70.

Mathieu, Ph., B. Beaufils and O. Brandouy (eds.) 2005. *Agent-Based Methods in Finance, Game Theory and their Applications*. Berlin: Springer.

Merton, Robert K. 1957. *Social Structure and Social Theory*, 2nd edn. London: The Free Press of Glencoe.

Miller, John H. and E. Scott Page. 2007. *Complex Adaptive Systems: an Introduction to Computational Models of Social Life*. Princeton University Press.

Morgan, S.L. 2005. *On the Edge of Commitment: Educational Attainment and Race in the United States*. Palo Alto: Stanford University Press.

Morgan, S.L. and Ch. Winship. 2007. *Counterfactuals and Causal Inference. Methods and Principles for Social Research*. Cambridge University Press.

Olson, James and Neal Roese. 2002. "Relative deprivation and counterfactual thinking," in Iain Walker and Heather Jean Smith (eds.), *Relative Deprivation. Specification, Development and Integration*. Cambridge University Press.

Pettigrew, Thomas. 2002. "Summing up: relative deprivation as a key social psychological concept," in Iain Walker and Heather Jean Smith (eds.), *Relative Deprivation. Specification, Development and Integration*. Cambridge University Press.

Pisati, M. 2007. "Unità della sociologia, unità della scienza," *Sociologica* 1.

Pujol, Josep M., Andreas Flache, Jordi Delgado and Ramon Sanguesa. 2005. "How can social networks ever become complex? Modelling the emergence of complex networks from local social exchanges," *Journal of Artificial Societies and Social Simulation* 8(4): 12.

Railsback, S.F., S.L. Lytinen and S.K. Jackson. 2006. "Agent-based simulation platforms: review and development recommendations," *Simulation* 82: 609–23.

Runciman, Walter Garrison. 1966. *Relative Deprivation and Social Justice: a Study of Attitudes to Social Inequality in Twentieth-century England*. London: Routledge and Kegan Paul.

Sanders, Lena. 2007. "Agent models in urban geography," in Frédéric Amblard and Denis Phan (eds.), *Agent-Based Models and Simulation for Human and Social Sciences*. Oxford: Bardwell Press.

Sawyer, Robert Keith. 2003. "Artificial societies. Multiagent systems and the micro-macro link in sociological theory," *Sociological Methods and Research* 31(3): 325–63.

Simon, H. 1979. "The meaning of causal ordering," in Robert K. Merton, James S. Coleman and Peter H. Rossi (eds.), *Qualitative and Quantitative Social Research*. New York: The Free Press, 65–81.

Stonedahl, F. and U. Wilensky. 2008. "NetLogo virus on a network model," Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.

Stouffer, Samuel A., Edward A. Suchman, Leland C. Devinney, Shirley Star and Robin M. Williams. 1965 [1949]. *The American Soldier. Vol. 1, Adjustment During Army Life*. New York: John Wiley & Sons.

Tesfatsion, L. and K.L. Judd (eds.) 2006. *Handbook of Computational Economics: Agent-Based Computational Economics*, vol. II. North-Holland: Elsevier.

Tisue, S. and Uri Wilensky. 2004a. "NetLogo: design and implementation of a multi-agent modeling environment," Evanston, Center for Connected Learning and Computer-Based Modeling, Northwestern University. Available at: http://ccl.northwestern.edu/papers.

2004b. "NetLogo: a simple environment for modeling complexity," Evanston, Center for Connected Learning and Computer-Based Modeling, Northwestern University. Available at: http://ccl.northwestern.edu/papers.

Tocqueville, Alexis de. 1955 [1856]. *L'ancien régime et la révolution*. Paris: Gallimard (here, English trans. Stuart Gilbert, New York: Doubleday, 1955).

Tyler, Tom R., Robert J. Boeckmann, Heather Jean Smith and Yuen J. Huo. 1997. *Social Justice in a Diverse Society*. Boulder: Westview Press.

Walker, H. and J. Smith (eds.) 2001. *Relative Deprivation. Specification, Development and Integration*. Cambridge University Press.

Wilensky, Uri. 1999. *NetLogo*. Evanston, Center for Connected Learning and Computer-Based Modeling, Northwestern University. Available at: http://ccl.northwestern.edu/netlogo.

2005. "NetLogo preferential attachment model," Evanston, Center for Connected Learning and Computer-Based Modeling, Northwestern University. Available at: http://ccl.northwestern.edu/netlogo/models/PreferentialAttachment.

Winship, Ch. and S.L. Morgan. 1999. "The estimation of causal effects from observational data," *Annual Review of Sociology* 25: 659–707.

Woodward, J. 2000. "What is a mechanism? A counterfactual account," *Philosophy of Science* 69(3): S366–S377.

Wolfers, Justin and Betsey Stevenson. 2008. "Economic growth and subjective well-being: reassessing the Easterlin paradox," IZA Discussion Papers, no. 3645.

Yamaguchi, Kazuo. 1998. "Rational-choice theories of anticipatory socialization and anticipatory nonsocialization," *Rationality and Society* 10: 163–99.

Index