Three concepts of causal mechanism in the social sciences

Tuukka Kaidesoja

Introduction

Presently, talk of causal mechanisms is popular among social researchers and methodologists. On some occasions, this term is used to develop new methodological ideas, but on other occasions, it is used to (re)describe and defend relatively old ways of conducting social research. The problem is that these different usages and related practices of causal analysis include very different assumptions about the nature of causal mechanisms, causation and the proper methods of causal analysis.

In this paper, I distinguish three concepts of causal mechanism that are significantly different. I am interested not only in semantic analysis of the meaning of these concepts but also in their practical implications for social research. The three concepts can be termed as (i) causal mechanisms as intervening variables, (ii) causal mechanisms as objective relations underlying the counterfactual dependence between variables in structural equation models and (iii) causal mechanisms as interaction structures of generative social processes. I specify the assumptions about causation and the most basic methodological ideas pertaining to each concept.

In addition, I argue that insofar as our aim is to provide a causal understanding about the processes of social interaction that include at least two interrelated social actors, the third concept of causal mechanism and the related methodological approach seem to be superior to the first two. In my view, the third concept also provides the most promising way of understanding the nature of social mechanisms. Nevertheless, I do not deny that the first two concepts and the associated methods of causal analysis are useful for some other epistemic purposes in social research.[1]

Causal mechanisms as intervening variables

Mahoney (2001, 578) writes that in sociology, "a causal mechanism is often understood as an intervening variable or set of intervening variables that explain why a correlation

[1] This paper elaborates and extends some ideas presented in my earlier paper on causal inference and modelling in sociology (Kaidesoja 2016). In that paper, I distinguished the three types of model-based causal inferences in sociology that assume different concepts of causation. Here, my focus is on the three concepts of causal mechanisms.

exists between an independent and dependent variable". In studies relying on this concept of causal mechanism, social researchers typically build and evaluate statistical models by means of analysing the correlations between the values of the variables in their data set that represent some population of interest. The idea is that the correlations between variables are accounted for by other correlations between variables and that causal inferences are made exclusively using correlational methods, such as cross tabulation and regression analysis. In this section, I specify this view of causal mechanisms and causal analysis as well as some of its problems.[2]

Advocates of this view maintain that a correlation between variables is a necessary condition for the establishment of causal relations. Nevertheless, it is generally accepted in the social sciences that the observed correlation between two variables as such is not a causal relation. The reason is that the observed correlation may turn out to be an accidental relation or it may have been produced by some confounding variable. To evaluate the probability that the correlation is an accidental relation, social researchers use various tests for significance that I do not consider here. In the case of non-accidental relations between two variables, still more is required for a relation to be causal than just the observed correlation between them. In textbooks on statistical analysis (in nonexperimental contexts), these extra requirements typically include at least the following:

> i. The values of the independent variable X for all units of analysis can be considered to be temporally prior to the values of the dependent variable Y;
>
> ii. There are no confounding variables $Z_1,\ldots,Z_n$ that have produced the observed correlation between the variables X and Y.[3]

According to a view that Goldthorpe (2000) calls *causation as robust dependence*, we have a causal relation between the correlating variables X and Y if these two conditions are met. In this view, then, causation is understood in terms of non-spurious correlations between observable variables. This view is often associated with the (probabilistic version of) Humean regularity theory of causation in which causal relations are reduced to the observable relations between variables or successive events.[4]

These two requirements for causal relations are best explicated by means of an example. In the simplest case, we may have an independent variable X that represents the number of years of formal education and a dependent variable Y that represents monthly income. By analysing our sample, we may observe that these continuous variables correlate in the sense that high levels of education are associated with high levels of income. In probabilistic terms, the high level of education may then be said to increase the probability of a high level of income for each unit of analysis. We may also reasonably assume

[2] Here, I ignore issues related to sampling and the representativeness of samples.

[3] It can also be added that the variables X and Y should not be conceptually related, in the sense that they constitute each other, and that we have some reasonable idea why they are correlated. Naturally, analysis of correlation between two variables first requires that there is some covariation in the values of the variables in our data set.

[4] Although David Hume's understanding of causation was more complex and nuanced than in the so-called regularity theory of causation, his views have inspired this theory. In addition, the regularity theory is more often implicitly presupposed than explicitly defended in the social scientific literature.

that requirement (i) is met in this example because people typically spend time in formal education before they start working regularly.

What about requirement (ii)? How can we know that the observed correlation between variables X and Y is not produced by some third variable Z that covaries with both X and Y? In social research using correlation methods, this question is typically answered by means of conditioning the potential confounding variables Z_i. In this operation, we analyse what occurs to the correlation between our variables X (representing the number of years of formal education) and Y (representing monthly income) when we hold the values of the relevant test variable Z (e.g., age or gender or ability) constant. If the correlation between X and Y does not disappear when the variable Z is conditioned (i.e., when the values of Z are held constant), then we can say that the correlation is robust (or non-spurious) with respect to the variable Z. If the correlation disappears in all classes of the variable Z (i.e., when Z takes different values), then we may contend that there is no direct statistical link between X and Y and that the variable Z that correlates with both X and Y statistically explains the observed correlation between X and Y. The implication is that the correlation between X and Y turns out to be spurious and that these variables are therefore not causally related. It is also possible that the observed correlation significantly changes (either decreases or increases) when we consider it in different classes of the variable Z. In such cases, the variable Z is said to moderate the statistical association between the variables X and Y. Finally, we may also find that there is some variable Z (e.g., representing occupation) in the causal path between the variables X and Y that (at least partly) mediates the relation between X and Y. These types of *intervening variables* (with their interpretations) are sometimes called causal mechanisms (e.g., Opp 2005; Morgan & Winship 2007, 224-226). This is the first concept of causal mechanism used in social research.

It is important to note that this concept of causal mechanism is horizontal in the sense that the intervening variable Z in the causal path between the variables X and Y represents the statistical properties of the population at the same level of aggregation as the variables X and Y (Kincaid 2012, 49). Therefore, this view does not require that we must provide any other empirical evidence for the existence of an intervening variable mechanism than that which can be produced by means of statistically analysing our data set. Naturally, in practice, social researchers provide theoretical interpretations of the intervening variables found through statistical analysis, but the point is that explanatory theories about social processes do not form an integral part of the causal analysis in the intervening variable view of causal mechanisms. These types of causal mechanisms, then, are not typically used to represent any theoretically specified causal process but, rather, statistical dependences between variables representing the distributions of properties in populations (cf. Goldthorpe 2001, 8-9; Opp 2005).

The conditioning of different "test variables" that may confound the observed correlation between two variables is sometimes called elaboration (e.g., Lazarsfeld 1957). Although the general idea of elaboration is both old and methodologically important, the above concept of causal mechanism is somewhat restricted in the sense that it is tied to the uses of correlation methods and the Humean regularity theory of causation. Nevertheless, it is important to recognize that the general methodological idea of elaboration can be usefully detached from correlational analysis, as exemplified in Ruonavaara's (2007;

also 2012, 3.1) view of "explanation by mechanisms". Ruonavaara (2007, 44; cf. 2012, 3.1) construes elaboration as a methodological operation of indicating how a connection between variables or factors is generated by means of outlining the causal process through which the connection is produced. Hence, I think that Ruonavaara's generalized view of elaboration can be best combined with the third concept of causal mechanism discussed in this paper.

In contrast, the intervening variable concept of causal mechanism assumes a Humean regularity theory of causation and, for this reason, does not require social researchers to go beyond the analysis of correlations between variables (see Goldthorpe 2001, 3-4; Mahoney 2001; Hedström 2005, Chapter 5). Instead, correlations are "explained" or "predicted" by means of other correlations, which means that this approach is both theoretically weak and vulnerable to accusations that one or more potentially confounding variables are omitted from the analysis. This account of causal mechanisms also lacks proper conceptual and methodological tools for representing social processes that consist of the interactions of social actors. Hence, to avoid these problems and restrictions, we need some additional idea that goes beyond correlational analysis to explicate why some correlations may be regarded as causal whereas others are merely accidental or spurious associations. The second view of causal mechanisms aims to provide such an additional idea while accepting that correlations between variables are necessary for making causal inferences.

Causal mechanisms as objective relations underlying the counterfactual dependencies between variables in structural equation models

The second causal mechanism concept is tied to the uses of structural equation models and the manipulationist account of causation. It is different from the intervening variable account of causal mechanisms because it rejects the Humean regularity theory of causation and denies that causal mechanisms are merely intervening variables (though it grants that causal relations may be *represented* by means of intervening variables). To specify the second concept of causal mechanism, we should examine the basic ideas of the *potential outcome framework of causal inference* that are assumed in this concept (e.g., Holland 1986; Sobel 1996; Härkönen 2004; Morgan & Winship 2007; Gangl 2010). This framework was systemized and generalized by Donald Rubin (1974); for this reason, it is occasionally called "the Rubin causal model". Since the potential outcome framework is based on experimentation practices in medical and agricultural sciences, its basic ideas can best be explicated by considering randomized controlled experiments.

In the simplest medical experiments of this kind, the experimenter randomly assigns the subjects into the test group and the control group. Then, she changes the value of independent variable X (e.g., a dichotomous variable representing the presence or absence of some medical treatment) in the test group (where X=1) and leaves it the same as before in the control group (where X=0). During the experiment, the experimenter either controls the potentially confounding variables Z_i in both groups or assumes that their effects are randomly distributed between the groups. At the end of the experiment, she measures the values of the dependent variable Y (e.g., representing some medical condition that the treatment is supposed to affect) in each subject and compares their

means with the means of the earlier values of Y that were measured before the treatment. If the experimenter can assume that the responses are on average the same in all subjects and that the treatments of different subjects are independent of each other, then she can legitimately infer that the difference in the averages of the measured changes with respect to the dependent variable Y between these two groups was caused by the manipulated change of the value of the independent variable X (i.e., whether the subjects got the treatment or not). In other words, once these conditions are met, an experimental set up like this enables the experimenter to control the effects of all variables Z_i that may potentially confound the dependence between the values of variables X and Y.

Now, taking its cue from these types of experimentation practices, the potential outcome framework assumes that we should be able (at least in principle) to manipulate the potential causes of the variation in the dependent variable if we want to make reliable causal inferences about the effects of these causes. This, in turn, requires that we have to know the causes whose effects we are going to analyse. This assumption is met in randomized controlled experiments because it is the experimenter who gives the treatment (e.g., new medicine) to the test group, though in double blind experiments, she does not know whether a given subject belongs to a test or a control group (whose members are given placebos). Analogically, in observational studies (i.e., studies where genuinely randomized and controlled experimentation is not possible), social researchers must know in advance the causes whose effects they are analysing (e.g., Morgan & Winship 2007; Knight & Winship 2013).

More generally, the potential outcome framework conceives causal relations in counterfactual terms. The implication is that if we consider the relation between the values of variables X and Y with respect to a particular unit of analysis, then we could in principle determine whether the measured variables are causally related by comparing how the value of the dependent variable Y with respect to the unit of analysis would have changed (or not) between two hypothetical situations: In the first situation, the unit is exposed to the treatment (i.e., it is assigned to the test group). In the second situation, the unit is not exposed to the treatment (i.e., it is assigned to the control group). The causal effect of the dichotomous independent variable X can then be written as y_t-y_c, where y_t is the value of variable Y in the case of treatment (x=1) and y_c is the value of variable Y in the case of non-treatment (x=0) of the unit. Since it is not possible for the same unit of analysis (e.g., a person) to be treated and not-treated (or to belong to the test group and the control group) simultaneously, this characterization of causation is counterfactual. In the potential outcome framework, randomized controlled experiments are then understood as *approximations* of these types of counterfactual situations since they enable us to estimate *the average effects of the cause variable X.*

It is important to recognize that we can only observe realized outcomes (e.g., the effects when the unit is treated), whereas making causal inferences, according to this view, ultimately requires that we compare the realized outcomes with the unrealized outcomes (e.g., the effects when the same unit is not treated), which by definition are unobservable. The so-called statistical solution to this "fundamental problem of causal inference" is to make causal inferences by means of focusing on the average effects of interventions (of some kind) in populations rather than by analysing singular causal processes in particular cases (e.g., Holland 1986; also Härkönen 2004, 57-58). To work, this solution requires

that certain conditions are met (e.g., units must be assigned to test and control groups randomly and the response of a unit cannot be affected by the treatment or non-treatment of other units). Hence, the potential outcome framework construes causal inference as "an effort to use observable data as a valid substitute to the unobservable (counterfactual) outcome information in order to estimate the causal effect of interest" (Gangl 2010, 25). In social scientific contexts where randomized controlled experiments are not typically possible, the possibility of making causal inferences is tied to the research designs and statistical methods that are used to imitate the randomized controlled experiments as much as possible (e.g., Morgan & Winship 2007).

Although the potential outcome framework requires that the cause variables to be manipulable in principle, manipulations of them do not have to be produced by social researchers but may also include "naturally occurring experiments" in populations (e.g., changes in social policy in a specific population). Nevertheless, the manipulation criterion gives an additional criterion for analysing the causal relations between variables that goes beyond empirical regularities. The core idea is that only the correlations between variables in statistical models in which the manipulation of the values of the cause variable would change the values of the effect variable(s) are understood as being causal (e.g., Woodward 2003, 15, 319-321). For example, if we would raise the average level of education of a certain population of individuals (e.g., by educating them), then doing so would increase the average income of the population if these positively correlated variables were also causally related in our statistical model. In manipulationist theories of causation, this type of manipulatability is considered not only a criterion for identifying causal relations but also a defining feature of all causal relations (e.g., ibid., Chapter 2). The potential outcome framework and manipulationist theory of causation can also be combined with either a deterministic or a probabilistic understanding of causal relations (e.g., ibid.).

More recently, the above views of causation and causal inference have been applied and adapted to the causal interpretation of structural equation models (e.g., Pearl 2000; Woodward 2003; Morgan & Winship 2007; Knight & Winship 2013). Roughly speaking, these types of models consist of many simultaneous regression equations that aim to account for causal relations between many variables that, together, form a causal system of interest. In building structural equation models, social scientists select suitable variables and specify the structural constraints of the model (i.e., the hypothesized relations between variables) on the basis of their antecedent theoretical assumptions and then test the structural equation model by using suitable data sets and estimation techniques. Because they utilize many equations to specify the relations among many variables, structural equation models enable social researchers to study not only the direct influences of many variables but also the indirect, net and latent influences. Hence, these types of models can be used to represent more complex causal systems compared to regression models.

Presently, structural equation models are often represented in the form of directed acyclic graphs (DAGs) that connect many variables and, thereby, specify the causal paths through which variables are assumed to affect each other and the relative weights of different relations between variables (e.g., Pearl 2000; Morgan & Winship 2007; Kincaid 2012; Knight & Winship 2013). Without going into the details of this method here, it suffices to indicate that a causal relation (i.e., the relation of counterfactual dependency) between two variable nodes that are connected by an arrow in the directed acyclic graph

is represented by the single equation in the structural equation model. Now, the second concept of causal mechanism refers to the objective relations underlying the counterfactual dependence between the values of variables in the single equation of the structural equation model that can be represented as an arrow in the directed acyclic graph (e.g., Woodward 2003, 48-49; cf. Woodward 2002; Knight & Winship 2013).

Although it is not possible to do full justice to these complex and ongoing discussions here, I highlight two important features of Woodward's (2003, Chapter 7; cf. Pearl 2000) causal interpretation of structural equation models since doing so enables me to further specify the second causal mechanism concept used in social research. According to Woodward (2003; 2013), the key features of the relations of counterfactual dependence between variables connected by the causal mechanism are that these relations are both invariant and modular.

They are invariant in the sense that the equations related to the specific causal paths in the true structural equation model are assumed to describe how a manipulation of the values of the independent variable would change the values of the dependent variable of the equation (e.g., Woodward 2003, 319). This is Woodward's (ibid. 15) more general characterization of the concept of invariance:

> A generalization G (relating, say, changes in the value of X to changes in the value of Y) is invariant if G would continue to hold under some intervention that changes the value of X such that, according to G, the value of Y would change - "continue to hold" in the sense that G correctly describes how the value of Y would change under this intervention.

Thus, the idea is that this account of causal mechanisms is not restricted to the representation of the actual causes of the variation in the dependent variable of the equation but also provides information about (a certain range of) counterfactual situations in which the values of the independent variable would have been different. It is then clear that this view is rooted in the potential outcome framework discussed above.

In addition, Woodward (ibid., 48; also Knight & Winship 2013, 282) assumes that the relations of counterfactual dependency between variables in a single equation in the structural equation model are assumed to be modular in the sense that "it is possible to disrupt or replace (the relationship represented by) any one of the equations in the system [e.g., structural equation model –T.K.] by an intervention on (the magnitude corresponding) the dependent variable in that equation, without disrupting any of the other equations". In other words, this view assumes that the causal mechanisms underlying these equations are isolated from each other so that they can be manipulated (at least to some extent) without affecting the other mechanisms in the system.

Insofar as I can observe, the invariance and modularity conditions (or closely similar conditions) related to this account of causal mechanisms are assumed by social researchers who rely on the potential outcome framework and the manipulationist theory of causation (e.g., Morgan & Winship 2007; Knight & Winship 2013; cf. Pearl 2000). One implication of this view of causal mechanisms is that the actual systems that can be successfully modeled from this perspective must resemble in the relevant respects mechanical machines that can be decomposed into distinct parts whose interconnections are relatively stable and independent of each other (e.g., Woodward 2013; Knight & Winship 2013, 282).

Although the second causal mechanism concept seems to work fairly well in the contexts where it is possible to conduct randomized controlled experiments (e.g., in agricultural research and the biomedical sciences), it unclear whether (or to what extent) the invariance and modularity conditions pertaining to the causal interpretation of structural equation models are suitable for the purposes of developing and testing causal models about *social processes that involve intentional actions and social interactions* (cf. Goldthorpe 2001; Hedström 2005, Chapter 5; Cartwright 2007, Part II; Gangl 2010, 40-41; Kincaid 2012). It seems to me that many processes of social interaction are impossible to decompose into distinct causal mechanisms in the above sense since the intentional actions of individuals are often dependent on specific institutional contexts as well as interdependent and complexly intertwined. This also means that, in such quasi-experimental settings that are utilized in social research, the knowledge of subjects about whether they belong to the "test group" or the "control group" may have far-reaching consequences on the effects of the "treatment" (e.g., some policy intervention or observed local social change) since this knowledge and the subjects' observations of each other's actions tend to affect how people respond to the treatment (or non-treatment) of interest (e.g., Goldthorpe 2001, 7-8). Hence, in the context of quasi-experimental research, social researchers can seldom rule out the possibility that relevant policy interventions may function (more or less) as self-fulfilling or self-destroying prophecies. For these reasons, it is unlikely that processes that involve these types of social interactions can be fruitfully modelled by using structural equation models that fulfil the invariance and modularity conditions. In addition, since it downplays the role of intentional actions of subjects who are the objects of treatments and other manipulations (see ibid. 7), this view of causal mechanisms and the related methodology of causal modelling seem to encourage a type of social engineering approach to society that may be considered suspicious among social researchers interested in critical theory and the analysis of power relations and structures.

It is also important to note that this view of causal mechanisms is ontologically flat (or horizontal) in the sense that the variables that are used to represent causal systems tend to be at "the same level of organization" or "aggregation" (Andersen 2014, 286; also Kincaid 2012). In this respect, this view resembles the intervening variable account of mechanisms rather than the third concept of causal mechanism discussed in the next section. Finally, the requirement that the cause variables should be always manipulable (at least in principle) poses restrictions on the type of counterfactual dependencies between variables that can be modelled by using causally interpreted structural equation models. The reason is that it denies that the so-called intrinsic variables (i.e., the variables that are part of the constitution of the units of analysis) can be among the possible causes because they cannot take different values (cf. Goldthorpe 2001, 6). Additionally, the necessary causal conditions for a phenomenon (i.e., the conditions that are always present when the phenomenon is present) and the causal processes involving feedback loops seem to pose problems for this account of causal mechanisms.

Insofar as these observations and considerations are correct, the above notion of causal mechanism seems to be very restricted for social research that aims to provide an explanatory understanding of processes involving social interactions. Thus, I doubt that the second causal mechanism concept combined with structural equation modelling can serve as a general model of causal analysis in the social sciences, though I do not deny its

usefulness for specific epistemic purposes, such as those pertaining to the experimental and quasi-experimental research designs that are not undermined by the abovementioned problems.

Causal mechanisms as interaction structures of generative social processes

Unlike the first two concepts, the third concept of causal mechanism used in the social sciences focuses on the social interactions of social actors rather than on the dependencies between variables in statistical models. It is also connected to the generative (or a realist) view of causation in which causal relations are analysed in terms of the causal powers, capacities and tendencies of entities and structures (e.g., Harré 1970; Harré & Madden 1975; Bhaskar 1979; Sayer 1984; Cartwright 1989; Little 1991; Manicas 2006) and the critique of the traditional ways of using statistical methods in causal analysis (e.g., Sayer 1984; Elster 1989; Little 1991; Bunge 1997; Hedström & Swedberg 1998). I begin by briefly describing the main points of the latter critique. Then, I consider how causal mechanisms are understood is interpreted in the movement of analytical sociology because analytical sociologists have provided the most detailed account of the third concept of causal mechanism and mechanism-based explanations in the social sciences.

In this view of causal mechanisms and mechanism-based explanations, the traditional approach to causal analysis in sociology is considered problematic because of the sociological implausibility of the causal models developed in this tradition (e.g., Hedström & Swedberg 1998; Sørensen 1998; Hedström 2005, 104; Goldthorpe 2001). The critics of the traditional approach often have in mind not only the correlation methods of causal analysis that pertain to the first concept of causal mechanism discussed above but also the statistical methods that are based on the potential outcome framework. Hence, advocates of the third causal mechanism concept typically reject both of the concept of causal mechanism discussed above (cf. Morgan & Winship 2007, Chapter 8).

Accordingly, this account of causal mechanisms shifts the interest from the analysis of the statistical distributions of properties and dependencies between variables to the social processes and social structures that are assumed to generate (or underlie) the statistical regularities that are observed by social researchers using statistical methods. This view underlies, for example, Hedström's (2005, 232) statement that "the core idea behind the mechanism approach is that we explain not by evoking universal laws, or by identifying statistically relevant factors, but by specifying mechanisms that show how phenomena are brought about". Furthermore, Hedström (ibid., 105) indicates that, in the traditional causal modelling using survey data, "theoretical statements have become synonymous with hypotheses about relationships between variables, and variables have replaced actors as the active subjects with causal powers". He also notes that randomized sampling, often used in survey research, implies that "individuals are uprooted from their social environments" (ibid., 2005, 109), meaning that the survey data about individuals who are randomly sampled from the larger population do not enable social researchers to analyse social interactions over time. As a result, according to Hedström (ibid. 106-107), the traditional statistical approaches to causal modelling either ignore the social interactions of individuals or make highly implausible assumptions about them (also see

Hedström & Swedberg 1998; Sørensen 1998; Goldthorpe 2001).

For these types of reasons, advocates of the third concept of causal mechanism emphasize that non-statistical methods of causal analysis and non-statistical evidence should play a major role in the production and evaluation of theories concerning causal mechanisms in social research. Although most of them grant that statistical methods are indispensable in establishing social phenomena that are worth explaining and in empirically evaluating our theoretical models concerning causal mechanisms, a number of non-statistical methods of causal analysis have been proposed for the purposes of developing and testing theories about processes of social interaction. They include narrative analysis (e.g., Manicas 2006, Chapter 5; Ruonavaara 2006), process tracing (e.g., Waldner 2012; Beach and Perdersen 2013), comparative process tracing (e.g., Bengtson & Ruonavaara 2017), and agent-based simulation (e.g., Hedström 2005, Chapter 4; Hedström & Ylikoski 2010, 62-64). Hence, the third concept of causal mechanism assumes a more pluralistic view of the proper methods of causal analysis compared to the first two concepts.

Although the promoters of these ideas share some common objects of criticism and some basic methodological and epistemological assumptions, they do not agree on the proper definition of causal mechanisms (for reviews of different accounts of causal and social mechanisms in the social sciences, see Mahoney 2001; Maynzt 2004; Hedström & Ylikoski 2010). Instead of discussing the details and relations between different definitions, I here focus on a specific interpretation of causal mechanisms and the related approach to mechanism-based explanations that have been developed in analytical sociology (e.g., Hedström 2005; Hedström & Bearman 2009).[5]

In his important textbook on the basic principles of analytical sociology, Hedström (2005, 1) writes that analytical sociology "seeks to explain complex social processes by carefully dissecting them and then bringing into focus their most important constituent components". This view suggests a "micro-foundationalist" research strategy according to which the macro-level outcomes of social processes are explained by referring to the micro-level mechanisms that are assumed to generate these outcomes (see Little 1991, Chapter 9). In other words, analytical sociologists aim to deliver an explanatory understanding of social macro-phenomena by developing precise, abstract and analytically realist explanatory theories and models that refer to psychological mechanisms and social mechanisms (Hedström 2005, Chapter 1).[6]They also often assume that social mechanisms are composed of interacting human individuals (e.g., ibid. 26-30; 34; 153-154; also Elster 1989; Hedström & Swedberg 1998) and that these types of mechanisms are relatively general (or portable) in the sense that they operate in recurrent social processes in different contexts rather than in a single process only. Nevertheless, Hedström (2005, 108) suggests that social mechanisms are better understood in terms of *causal tendencies* rather

[5]Elsewhere, I have compared analytical sociologists' views of causal and social mechanisms to the notions of explanatory mechanism and mechanism-based explanation that have been developed in the critical realist movement (see Kaidesoja 2013aa; also see Ruonavaara 2007; 2012). Here, it suffices to say that critical realists tend to connect social mechanisms more tightly to social structures than analytical sociologists although it is also possible to develop mediating positions between these views (e.g., Ruonavaara 2012; Kaidesoja 2013a, Chapter 6; 2013bb).

[6]In this paper, I do not discuss psychological mechanisms, though I agree with analytical sociologists that, in addition to social mechanisms, psychological mechanisms may be referred to in explanations of social phenomena (e.g., Elster 1989; Hedström 2005).

than in terms of deterministic processes since other operative mechanisms may modify the effects of the activated social mechanism of interest in any concrete social processes whose outcome is generated by multiple interacting mechanisms (cf. Bhaskar 1979; Sayer 1984; Goldthorpe 2001, 12).

More specifically, Hedström (2005, 25), drawing on Machamer, Darden and Craver's (2000) influential paper on causal mechanisms in the sciences, writes that:

> [causal – T.K.] mechanisms can be said to consist of entities (with their properties) and the activities that these entities engage in, either by themselves or in concert with other entities. These activities bring about change, and the type of change brought about depends upon the properties of the entities and the way in which they are linked to one another.

This is an ontic view of causal mechanisms according to which concrete instances of mechanisms are thought to exist independently of the scientific theories and models concerning them (e.g., Hedström 2005, 14). Scientists, in turn, can use their theories and models to represent these mechanisms more or less abstractly and accurately. Hedström (ibid., Chapter 3) further assumes that instances of *social mechanisms* consist of social interactions of interrelated and socially situated human individuals in which they affect each other's beliefs, desires, emotions and opportunities. This view of social mechanisms is rooted in the tradition of methodological (or structural) individualism and is shared by many other analytical sociologists (e.g., Hedström & Swedberg1998; also Elster 1989; Hedström & Bearman 2009).

Since analytical sociologists assume that social mechanisms are composed of individual actors and their social interactions, it is impossible to describe social mechanism without some action theoretical assumptions (Ruonavaara 2012, 3.4-3.6). Although some social researchers adopt rational choice theory for that purpose, analytical sociologists often rely on the looser DBO theory of action in which the actions and interactions of social actors are understood in terms of their desires, beliefs and opportunities (Hedström 2005, 38-42; cf. Kaidesoja 2012). Regardless, to borrow Ruonavaara's (2012) useful terminology, some sort of *agent-image* is needed to develop theories and models about social mechanisms. I have elsewhere suggested that the explanatory interests of analytical sociologists would be best served if they applied different action theoretical assumptions (i.e., agent images) for different explanatory purposes rather than rely on some action theory that is considered universal (e.g., the DBO theory of action or rational choice theory) (Kaidesoja 2012).

In analytical sociology, theoretical models about social mechanisms are assumed to open up black boxes by detailing "the causal cogs and wheels" (Hedström & Ylikoski 2010, 54) of the social processes through which certain types of social phenomena are regularly brought about. These models often trace such interdependencies in the interactions of social actors (e.g., thresholds for action that depend on the actors' perceptions of the actions of other actors as well as the feedback and feedforward loops occurring through social interactions) that are difficult to represent with structural equation models that assume the modularity condition discussed above. Following Merton's (1968, Chapter 2) ideas, it is also emphasized that sufficiently clear and precise middle-range theories about social mechanisms should focus on only the limited aspects of concrete social processes and that it is futile to aim to provide a comprehensive explanation of any social phenomenon

(e.g., Hedström & Ylikoski 2010, 61-62). In addition, middle-range theories about social mechanisms perform an important role in causal analysis because our knowledge about the social mechanisms that connect the hypothesized causes and effects supports causal inferences. Conversely, the lack of knowledge about any plausible social mechanism that connects the alleged causes and effects gives us a reason to doubt whether they are causally related at all (Hedström & Ylikoski 2010, 54).

Although the borders of the analytical sociology movement are somewhat vague, examples of commonly cited theoretical models (or middle-range theories) concerning social mechanisms include the following:

- Self-fulfilling prophecies (Robert Merton)
- The Matthew effect and the theory of cumulative advantage (Robert Merton)
- Diffusion mechanisms in social networks (e.g., James Coleman)
- Social segregation mechanisms (e.g., Thomas Schelling)
- Threshold mechanisms of social influence and collective action (e.g., Mark Granovetter)

Although this list of theoretical models is far from comprehensive, all of these models refer to social mechanisms that include social interactions. All of these models have also generated important theoretical discussions and empirical applications. Therefore, they may be said to belong to the theoretical toolbox of analytical sociologists.

As suggested above, this view of social mechanisms is not tied to any specific method of causal analysis. Instead, analytical sociologists maintain that there are many different methods that can be used to provide evidence for the existence and operations of social mechanisms in different contexts. For this reason, analytical sociologists "make a clear distinction between statistical analysis and sociological explanations" (Hedström 2005, 113; also Hedström & Swedberg 1998; Manzo 2010). Hence, although analytical sociologists do not reject the uses of statistical methods, they emphasize that other types of methods are also needed in causal analysis and explanatory social research.

I think that one of the strengths of analytical sociology is that it has paid attention to the importance of substantial middle-range theories and models concerning social mechanisms that cannot be reduced to statistical models. As Ruonavaara (2012, 5.2.) suggests, analytical sociologists have mostly been interested in modelling social mechanisms in which relatively uncoordinated but interdependent intentional actions and interactions of many interrelated individuals produce macro-level outcomes that are unintended by any of these actors. I think that this view is correct, but I see no reason to deny that the conceptual tools developed in this movement may also be used to develop theoretical models concerning social mechanisms that consist of socially coordinated interactions of individuals and whose outcomes are collectively intended by the relevant actors.

Although it cannot be denied that the movement of analytical sociology has brought new ideas to the methodological debates on causal explanations and social mechanisms, a number of critiques have also been raised against the concept of causal mechanism. For example, it has been suggested that the idea of a generative mechanism as such does not

provide us with a method for verifying theoretical models concerning causal mechanisms and, for this reason, should be replaced with the second concept of causal mechanism discussed above connected to the method of structural equation modelling (e.g., Morgan & Winship 2007, 237; 240-242; also Knight & Winship 2013). Although this suggestion is presented as a supplement to rather than a direct critique of analytical sociology, I think that it actually requires relinquishing the third concept of causal mechanism in favour of the second concept (insofar as the above interpretations and analyses are sound). In addition, it has also been argued that analytical sociology is too individualistic and reductionist in approach and that it does not do full justice to the social mechanisms in which various institutions, organizations and cultural contexts play a crucial role (e.g., Mayntz 2004; Little 2012; Kaidesoja 2013b). Although it is not possible to go into the details of these critiques here, I briefly comment on both of them.

It is certainly true that analytical sociology does not provide a method for the verification of theories and models concerning causal mechanisms. Nevertheless, I doubt whether the strictly empiricist epistemology and the monistic approach to causal analysis assumed in the first suggestion/critique are suitable for social research. Although I think that social researchers should aim at developing theories and models that can be evaluated empirically (in the sense that empirical evidence may be used to support, reject or correct them), it seems to me that, if taken literally, the requirement that these theories and models should be verified (i.e., shown to be true on the basis of empirical evidence)[7] is not feasible for the purposes of social research (also see Goldthorpe 2001, 10). In my view, there are no good reasons to require that the strictly empiricist epistemology of this kind should be adopted in the social sciences, given that it is not universally accepted in other empirical sciences either. Furthermore, it seems to me that the methodologically monistic orientation towards causal analysis in terms of structural equation models in which this type of critique is rooted is too restricted for the purposes of social research. First, as argued above, it poses remarkable restrictions on causal mechanisms pertaining to social processes that can be modelled by using causally interpreted structural equation models. Second, it requires that models concerning causal mechanisms should always be tested by using research designs that imitate randomized controlled experiments in relevant respects, which, I think, is not a feasible requirement in the social sciences. This is not to deny, however, that experimental and quasi-experimental methods, among other methods of causal analysis, can be used in social research for limited epistemic purposes. Third, the monistic view of causal analysis can also lead to the illegitimate and counterproductive rejection of the developing non-statistical methods of causal analysis that were noted above.

Hence, I think that it is important to recognize that the methodologically pluralist orientation that is connected to the third concept of causal mechanism enables the empirical testing of theories and models about social mechanisms by using different types of data and methods of empirical analysis. This, in turn, allows social researchers to evaluate

[7] I am not certain whether this literal interpretation of the concept of verification is intended by Morgan and Winship (2007, 237) when they write about how explanatory depth is "best secured when it is verified in empirical analysis grounded on the counterfactual model" that is connected to the second concept of causal mechanism discussed above. Nevertheless, this is a conceivable critique of the third concept of causal mechanism.

competing theories about social mechanisms that are hypothesized to have generated a specific outcome in terms of whether the operations of the postulated mechanisms can be traced by using different methods. If the causal mechanism postulated by the theory is robust in the sense that the theory is confirmed with evidence produced by using different methods and data sets that are independent of each other, then our confidence in the theory increases, particularly if the mechanisms postulated by competing theories fail to show this sort of robustness (see Wimsatt 2007, Chapter 4). Nevertheless, detailed issues about the empirical evaluation of theories concerning causal mechanisms and mechanism-based explanations go beyond this paper.

Since I have been among the critics who have raised the second critique, I agree that it has a grain of truth. At least in their early programmatic texts, analytical sociologists have argued for methodological individualism and denied the existence of social entities and macro-mechanisms (e.g., Hedström & Swedberg 1998; Hedström 2005; also Elster 1989). Nevertheless, I tend to think that, though they were important in the initial formation of this movement, these individualist assumptions should not be viewed as the fundamental ingredients of analytical sociology, and they have been downplayed in more recent programmatic writings by analytical sociologists (e.g., Hedström & Ylikoski 2010; Manzo 2010).

As I have elsewhere argued in detail, I also think that the notion of a generative mechanism and the related generative theory of causation are perfectly compatible with a multilevel account of social mechanisms according to which there can be "social macro-mechanisms" (i.e., social mechanisms that are composed of institutionally embedded collective actors) in addition to "social micro-mechanisms" (i.e., social mechanisms that are composed of socially situated human individuals) (Kaidesoja 2013b; Kaidesoja & Kauppinen 2014). For example, in cases where a social researcher can assume that the relevant actors consist of interacting and relatively stable organizations with specific capacities (e.g., a capacity for making and implementing collective decisions), I think that explanations in terms of macro-mechanisms that are composed of organizations would be compatible with the third view of causal mechanisms. Although, naturally, it is always possible and sometimes useful to zoom in to study the causal mechanisms that underlie the capacities and actions of specific organizations, the point is that there are explanatory questions that are best answered in terms of macro-mechanisms composed of interrelated and interacting organizations in a certain institutional framework (for some examples, see Kaidesoja 2013b; Kaidesoja & Kauppinen 2014). How these types of important macro-mechanisms turn out in social research is an empirical question.

Conclusion

I have distinguished three different concepts of causal mechanism used in social research. In the first concept, causal mechanisms are interpreted as intervening variables. This view assumes a Humean regularity theory of causation and is connected with the traditional uses of correlation methods in social research that are often based on the idea of elaboration. The second concept construes causal mechanisms as being objective relations that underlie the counterfactual dependencies between the values of variables that

are represented by the regression equations connecting the cause and effect variables in a structural equation model. This concept of causal mechanism includes an assumption that different equations in a structural equation model are invariant and modular, which poses very severe restrictions on the social processes that can be successfully modelled in this approach. The third concept assumes that the most important causal mechanisms studied in the social sciences are social mechanisms that consist of the interaction structures of generative social processes. Unlike the first two concepts, the third concept is not restricted to the uses of statistical methods but is more pluralistic with respect to the proper methods of causal analysis.

Insofar as my distinctions and interpretations are sound, there are (at least) three different concepts of causal mechanism used in social research that include incompatible assumptions on the nature of causation and the proper methods of causal analysis. I think that they should be kept separate to avoid unnecessary conceptual confusions. I have also suggested some reasons to think that the third concept of causal mechanism and the related pluralistic approach to causal analysis are the most promising if our aim is to produce an explanatory understanding of processes of social interaction. Although I have focused here on explanatory social research, I do not deny that we also need descriptive social research for many epistemic and practical purposes. I also want to leave it open here whether there are other types of explanations of social phenomena in addition to causal explanations.

References

Andersen, H. (2014). A field guide to mechanisms: Part II. *Philosophy Compass*, 9(4):284–293.

Beach, D. and Pedersen, R. B. (2013). *Process-tracing methods: foundations and guidelines*. The University of Michigan Press: Ann Arbor.

Bengtsson, B. and Ruonavaara, H. (2017). Comparative process tracing: Making historical comparison structured and focused. *Philosophy of the Social Sciences*, 47(1):44–66.

Bhaskar, R. (1979). *The possibility of naturalism: a philosophical critique of the contemporary human sciences*. Harvester Press, Brighton.

Bunge, M. (1997). Mechanism and explanation. *Philosophy of the Social Sciences*, 27(4):410–465.

Cartwright, N. (1989). *Nature's capacities and their measurement*. Oxford University Press, Oxford.

Cartwright, N. (2007). *Hunting causes and using them: approaches in philosophy and economics*. Cambridge and New York: Cambridge University Press.

Elster, J. (1989). *Nuts and bolts for the social sciences*. Cambridge University Press, Cambridge; New York.

Gangl, M. (2010). Causal inference in sociological research. *Annual Review of Sociology*, 36:21–48.

Goldthorpe, J. H. (2000). *On sociology: numbers, narratives, and the integration of research and theory*. Oxford University Press, Oxford.

Goldthorpe, J. H. (2001). Causation, statistics, and sociology. *European Sociologigal Rewiew*, 17:1–20.

Härkönen, J. (2004). Kausaalinen päättely ja sosiologinen tutkimus. In Räsänen, P., Ruonavaara, H., and Kantola, I., editors, *Kiistoja ja dilemmoja: Sosiologisen keskustelun vastakkainasetteluja*. Kirja-Aurora, Turku.

Harré, R. (1970). *The principles of scientific thinking*. Macmillan, London.

Harré, R. and Madden, E. H. (1975). *Causal powers: a theory of natural necessity*. Blackwell, Oxford.

Hedström, P. (2005). *Dissecting the social: On the principles of analytical sociology*. Cambridge University Press, Cambridge.

Hedström, P. and Bearman, P. (2009). *The Oxford handbook of analytical sociology*. Oxford University Press, Oxford.

Hedström, P. and Swedberg, R. (1998). Social mechanisms: an introductory essay. In Hedström, P. and Swedberg, R., editors, *Social Mechanisms: An Analytical Approach to Social Theory*. Cambridge University Press, Cambridge.

Hedström, P. and Ylikoski, P. (2010). Causal mechanisms in the social sciences. *Annual Review of Sociology*, 36:49–67.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81:945–960.

Kaidesoja, T. (2012). The DBO theory of action and distributed cognition. *Social Science Information*, 51(3):311.

Kaidesoja, T. (2013a). *Naturalizing critical realist social ontology*. Routledge, London.

Kaidesoja, T. (2013b). Overcoming the biases of microfoundationalism social mechanisms and collective agents. *Philosophy of the Social Sciences*, 43(3):301–322.

Kaidesoja, T. (2016). Causal inference and modeling. In McIntyre, I. and Rosenberg, A., editors, *The Routledge Companion to Philosophy of Social Science*. Routledge, New York.

Kaidesoja, T. and Kauppinen, I. (2014). How to explain academic capitalism: A mechanism-based approach. In Cantwell, I. and Kauppinen, I., editors, *Academic Capitalism in the Age of Globalization*. Johns Hopkins University Press, Baltimore.

Kincaid, H. (2012). Mechanisms, causal modeling, and the limitations of traditional multiple regression. In Kincaid, H., editor, *The Oxford Handbook of Philosophy of Social Science*. Oxford University Press, Oxford.

Knight, C. and Winship, C. (2013). The causal implications of mechanistic thinking: Identification using directed acyclic graphs (DAGs). In Morgan, S., editor, *Handbook of Causal Analysis for Social Research*. Springer, Dordrecht.

Lazarsfeld, P. (1957). Interpretation of statistical relations as a research operation. In Lazarsfeld, P. and Rosenberg, A., editors, *The Language of Social Research*. The Free Press, Glencoe.

Little, D. (1991). *Varieties of social explanation: An introduction to the philosophy of social science*. Westview Press, Boulder and Oxford.

Little, D. (2012). Analytical sociology and the rest of sociology. *Sociologica*, 12(1):1–47.

Machamer, P., Darden, L., and Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, 67(1):1–25.

Mahoney, J. (2001). Beyond correlational analysis: Recent innovations in theory and method. *Sociological Forum*, 16(3):575–593.

Manicas, P. T. (2006). *A realist philosophy of social science: explanation and understanding*. Cambridge University Press, Cambridge.

Manzo, G. (2010). Analytical sociology and its critics. *European Journal of Sociology; Cambridge*, 51(1):129–170.

Mayntz, R. (2004). Mechanisms in the analysis of social macro-phenomena. *Philosophy of the Social Sciences*, 34(2):237–259.

Merton, R. K. (1968). *Social theory and social structure*. The Free Press, New York, 2 edition.

Morgan, S. L. and Winship, C. (2007). *Counterfactuals and causal inference: methods and principles for social research*. Cambridge University Press, Cambridge.

Opp, K.-D. (2005). Explanations by mechanisms in the social sciences. Problems, advantages and alternatives. *Mind & Society*, 4(2):163–178.

Pearl, I. (2000). *Causality: models, reasoning, and inference*. Cambridge University Press, Cambridge.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, 5(66):688–701.

Ruonavaara, H. (2006). Historian polut ja teorian kartta - eli miten tutkia tapahtumaketjuja sosiologisesti. In Saari, J., editor, *Historiallinen käänne: Johdatus pitkän aikavälin historian tutkimukseen*. Gaudeamus, Helsinki.

Ruonavaara, H. (2007). Mekanismeilla selittäminen. *Sosiologia*, 45(1):37–51.

Ruonavaara, H. (2012). Deconstructing explanation by mechanism. *Sociological Research Online S*, 17(2):1–10.

Sayer, A. (1984). *Method in social science: a realist approach*. Routledge, London, 2 edition.

Sobel, M. E. (1996). An introduction to causal inference. *Sociological Methods and Research*, 24(3):353.

Sørensen, A. (1998). Theoretical mechanisms and the empirical study of social processes. In Hedstrom, P. and Swedberg, R., editors, *Social mechanisms: An analytical approach to social theory*. Cambridge University Press, Cambridge.

Waldner, D. (2012). Process tracing and social mechanisms,. In Kincaid, H., editor, *The Oxford Handbook of Philosophy of Social Science*. Oxford University Press, Oxford.

Wimsatt, W. C. (2007). *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Harvard University Press, Cambridge, MA.

Woodward, J. (2002). What is a mechanism? A counterfactual account. *Philosophy of Science*, 69(3):366–377.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press, Oxford.

Woodward, J. (2013). Mechanistic explanation: Its scope and limits. *Proceedings of the Aristotelian Society. Supplementary Volume*, 87(1):39–65.