9 Causal explanation through social mechanisms

In order to provide causal explanations for established population regularities, causal processes, or mechanisms, must be hypothesized in terms of individual action and interaction that meet two requirements: they should be in principle adequate to generate the regularities in question and their actual operation should be open to empirical test. Advantage lies with mechanisms explicitly specified in terms of action that is in some sense rational.

Until well into the twentieth century, it was the standard view that causal explanations in science were arrived at by showing how observed phenomena followed from the operation of some general 'covering' law of a deterministic kind (see e.g. Hempel, 1965). And this view does indeed in various quarters persist.[1] However, with the probabilistic revolution, as discussed in Chapter 1, the idea of causal explanation as being dependent on the existence of deterministic laws was called into question, and in the more recent past a significantly different idea of the nature of such explanation has emerged and gained in acceptance. This is, in brief, the idea that causal explanations entail the spelling out, as fully as possible, of just *how* – through what continuous space–time processes or *mechanisms* – a supposed cause actually produces its effect (for extensive discussion from the standpoint of the philosophy of science, see Illari, Russo and Williamson, 2011).

Proponents of what could be called 'mechanism-based causal explanation' recognise that, from field to field, the nature of the

[1] For example, those who maintain from within sociology that it cannot become a science often attach great importance to its failure to produce general laws. But what is thus indicated is a limited understanding both of developments within the philosophy of science and of actual practice across the sciences – or, at all events, an undue preoccupation with classical physics. On the way in which the biological sciences offer far more instructive parallels for sociology as regards models of explanation, and more generally, see Lieberson and Lynn (2002).

mechanisms that will need to be envisaged and the ways in which they will be specified, in terms of the entities involved and their causal capacities, will vary widely. It may be recalled that Neyman – who can be regarded as an early adherent of the idea of mechanism-based explanation – emphasised that the mechanisms to be invoked in a population science in order to explain aggregate-level probabilistic regularities would be ones that, rather than being deterministic and applying to every individual case, themselves 'incorporated chance'. And in sociology understood as a population science, it would further seem clear that the key entities of such mechanisms must be individuals and that causal capacity must be taken to lie in the action of individuals and, ultimately, in the degree of autonomy that, through the possibility of informed choice, such action possesses (see Chapter 3). From this standpoint, then, the critiques previously noted of variable sociology in general and of regression analysis in particular for their neglect of the individual action underlying statistically established regularities can be taken to imply 'a plea for mechanisms' (Elster, 1998). What the movement in favour of mechanism-based explanation in sociology is essentially in search of is, in the words of Hedström and Bearman (2009a: 5), a means of 'making intelligible the regularities being observed by specifying in detail how they were brought about' – or, in other words, making these regularities not only visible but transparent.

It is important here to emphasise that seeking causal explanations in sociology in terms of mechanisms does not in itself entail some further advance of a technical kind. In particular, it is not a matter, as seems sometimes to be supposed, of simply including more possible 'intervening' variables within a statistical analysis or drawing more complex causal-path diagrams or graphs. The crucial input has to be a sociological and theoretical one. More specifically, what is needed and what, I would argue, sociologists seeking mechanism-based explanations have in fact generally aimed to provide, are what might be called *generalised narratives of action and interaction* that underlie regularities that call for explanation. To be of (potential)

explanatory value, the mechanisms that are represented in these narratives need to possess two key features. First, they have to be (to use the term in a somewhat different sense to Max Weber) causally adequate: that is to say, it must be possible to show how, through individuals acting and interacting in the ways that are spelled out, the regularities of interest *could* be generated and sustained. Second, though, the narratives have to be ones expressed in such a form that the question of whether or not the mechanisms they specify *do in fact operate* in the way hypothesised is open to empirical examination, so that through further research the explanation offered can be either rejected or corroborated.

Some similarity may be seen between the formulation of narratives of the kind in question here and what proponents of qualitative case studies (e.g. Collier, Brady and Seawright, 2004; George and Bennett, 2005: ch. 7) refer to as 'causal process tracing'. But, as Bennett (2008: 704) has acknowledged, insofar as this approach is applied to the explanation of singular events rather than of established regularities, what is involved is the spelling out of quite specific causal sequences – or, in other words, historical explanation (see Chapter 4) – rather than the identification of mechanisms that recurrently operate. As Elster (1998: 45–9) has observed, while explanations in terms of mechanisms have less generality than do explanations in terms of covering laws, they still aim to have greater generality than narratives of a quite idiographic kind.

In the development of mechanism-based explanations in sociology, two different approaches can be identified. Through distinguishing and comparing these approaches, certain issues of major importance can be brought out concerning the actual practice of constructing explanations in sociology as a population science.

One approach is that pursued most explicitly by Elster (1989, 2007), but which is also generally favoured by adherents of what has become known as 'analytical sociology' (Hedström and Swedberg, 1998b; Hedström and Bearman, 2009b). In this case, the aim could be seen as that of creating a kind of *catalogue raisonné* of mechanisms

that operate in social life, ranging from the most elementary through to the most complex. What is then apparently envisaged is that sociologists confronted with an explanatory problem will be able to search this catalogue for mechanisms that would appear most likely to lead to a solution; or, to use the metaphor favoured by Elster (1989: 3) – and adopted by Hedström and Bearman – sociologists will be able to draw on the 'toolbox of mechanisms – nuts and bolts, cogs and wheels' that is made available.

The main advantage of this approach is that it opens up the possibility of theoretical integration and systematic development through the same or similar mechanisms being found to operate across a range of different substantive domains. And some success might in this regard be claimed – as, say, in the case of various social diffusion mechanisms, 'Matthew-effect' mechanisms of cumulative advantage and signalling mechanisms (for assessments, see Palloni, 1998, DiPrete and Eirich, 2006 and Gambetta, 2009, respectively).

However, the approach also has its dangers. Perhaps the most apparent is that it can give rise to a greater interest in mechanisms *per se* than in the extent of their explanatory potential: that is, beyond cases specially selected so as to best illustrate their application. Particular mechanisms and *what they might explain*, rather than established but non-transparent social regularities and *how they are to be explained*, become the foci of attention. In this way, then – and to return to the discussion of Chapter 8 – a concern becomes more evident with the effects of causes than with the causes of effects; or, one could say, attention centres simply on causal adequacy, in the sense previously indicated. However, as also indicated, while causal adequacy is necessary to a successful mechanism-based explanation, it is not sufficient. Evidence needs also to be provided that a hypothesised mechanism is that which does actually operate to produce the regularities that are under examination in any particular instance (Erikson, 1998).

A second approach to the development of mechanism-based explanations is then one which might be regarded as more congruent

with the idea of sociology as a population science. This is the approach followed by sociologists whose starting point is with probabilistic population regularities that have been established in some substantive field of research but that remain without a satisfactory explanation: that is to say, they remain opaque rather than transparent regularities – the ways in which they derive from individual action and interaction, under the conditions of action that prevail, are not well understood. The question to be faced is therefore clearly one of the causes of effects.

This second approach, it has to be said, is not as developed as the first. As noted in Chapter 1, those sociologists who do focus their attention on population regularities have achieved far more in describing these regularities than in explaining them, or, that is, in making them visible rather than making them transparent. It has also to be recognised that, where causal mechanisms are hypothesised in relation to specific regularities, they may take on a somewhat *ad hoc* character. However, it is also likely that in such cases more than one mechanism can be envisaged, and in this way the importance is underlined of empirical testing designed to determine the relative merits of the differing explanations that are on offer.

It could thus be regarded as the most important feature of the second approach to mechanism-based explanation that, in pursuing it, one further quite crucial issue for sociology as a population science is brought to the fore: namely, that of *how – through what forms of research – one can best determine the actual operation of causal mechanisms*: or, that is, of social processes that lie, to revert to Cox's (1992: 297) phrase, at 'an observational level that is deeper than that involved in the data under immediate analysis'. As regards the methods of data collection and analysis they entail, these forms of research do not have to be the same as, and may in fact need to be different from, those that are essential in enabling population regularities – the explananda – to be reliably and accurately established.

To illustrate the problems – and the possibilities – that arise here, I consider attempts at explaining regularities that have become

established in regard to inequalities of educational attainment among children of differing social backgrounds, and in particular of differing social class backgrounds. What has been shown through statistical analysis of survey data of various kinds is that these inequalities come about in two different ways, labelled as 'primary' and 'secondary' effects (Jackson, 2013). First, children from more advantaged class backgrounds on average perform better educationally than do children from less advantaged backgrounds: that is, in regard to grades, tests, examinations and so on; but, second, children from more advantaged backgrounds also tend to make more ambitious educational choices than do children from less advantaged backgrounds *even when level of previous performance is held constant*.

Advances in the understanding of primary effects have been made, and continue to be made, through the analysis of the complex interaction of sociocultural, economic and genetic influences at work; but secondary effects pose a different and also a more specifically sociological problem.[2] To seek to explain these effects simply by appealing to social class differences in values and norms relating to education is inadequate, since it is generally the case in modern societies that young people from all class backgrounds alike are steadily raising their levels of educational aspiration, participation and attainment, even while marked inequalities persist (see further Goldthorpe, 2007: vol. 2, chs 2–4). What is required is some narrative that, consistently with the individualistic paradigm, does not rely on unreflective and unconditional norm-following but takes account of individuals' ends, the constraints – non-normative as well as normative – under which they pursue these ends and the informed choices that they then make to pursue one course of action rather than another.

[2] In the case of primary effects, it is apparent that the causal mechanisms that operate are not only ones of sociological interest that can be expressed in terms, say, of the actions of parents relevant to their children's chances of educational success, as conditioned by the differing forms and levels of resources available to them, but also mechanisms that fall in the domains of epigenetics, neuroscience and developmental psychology.

A number of mechanisms on the lines in question have in fact been suggested. Certain authors (e.g. Esser, 1999; Becker, 2003) have adopted a rather standard 'expected utility' approach from micro-economics. Others, including Richard Breen and myself (Goldthorpe, 1996; Breen and Goldthorpe, 1997; cf. Erikson and Jonsson, 1996), have proposed mechanisms that are based on a more bounded and less demanding 'rationality of everyday life' (see Chapter 3). In what follows, I concentrate on what has become known as the Breen–Goldthorpe 'relative risk aversion' (RRA) theory, not so as to privilege my own work but simply because my concern is with how mechanism-based explanations in sociology are to be evaluated through further research and because the RRA theory has in fact been subject to far more, and more varied, empirical testing than others in the field.[3]

The basic claim of the RRA theory is that, when making educational choices with their futures in mind, young people, and their parents, will give priority to the avoidance of downward social mobility over the achievement of upward mobility.[4] However, while risk aversion can then be seen as equal, relative to social origins, *the actual risks involved* in educational choice will be unequal. For children of

[3] In Goldthorpe (2007: vol. 2, ch. 4), I review results from six different tests, and others could have been included: Holm and Jaeger (2008) refer to four further tests prior to making one of their own (see subsequent text), and I am aware of several others of still more recent date. I previously concluded that, although various difficulties with the RRA theory had been revealed and the need for refinements and further development indicated, it remains, in its essentials, 'alive'. And this is the position I would still adhere to. Interesting attempts at taking the theory further can be found in Breen and Yaish (2006) and Breen, van de Werfhorst and Jaeger (2014). I have myself (Goldthorpe, 2007: vol. 2, ch. 7) attempted to extend the theory to intergenerational class mobility and hope to continue this work in the light of further empirical research into social mobility in which I am currently involved.

[4] The theory could then be regarded as a special case of the more general 'prospect theory' propounded by Kahneman and Tversky (1979), according to which the slope of individuals' utility curves is steeper in the domain of losses than in the domain of gains. However, I would not myself wish to follow Kahneman (2011: 286) in supposing that a 'failure of rationality' is 'built into prospect theory' – or, therefore, into the RRA theory – simply because it violates the logic of choice inherent in expected utility theory (see Chapter 3, n. 7).

more advantaged social origins, there will be little to lose, in seeking to maintain their parents' position, by taking up all further educational opportunities that their previous performance makes available to them – and even if their chances of ultimate success may be doubtful. But for children of less advantaged social origins, educational choice will be more problematic. For, in their case, more ambitious choices that might end in failure could not only be in various ways costly in themselves but could also preclude less ambitious choices that, even if not offering great prospects for advancement, would at all events still effectively guard against downward mobility. Consequently, for these children to make a more ambitious choice, they would need to have a greater assurance of success – as would be indicated by a higher level of previous performance – than would their more advantaged counterparts.

If the mechanism spelled out by such a narrative were in operation then secondary effects in class inequalities in educational attainment would be generated through rather straightforward aggregation. The mechanism could, in other words, be regarded as causally adequate. But how can it be determined whether, or how far, it is in fact at work? At least three different research strategies can in this regard be identified – each, of course, requiring that, to revert to the argument of Chapter 5, relevant variables should be conceptualised and made operational in appropriate ways.

The first strategy could be described as that of direct observation. If causal mechanisms are understood as continuous space–time processes, then it should, in principle, be possible to obtain direct evidence of their operation wherever this is going on, and intensive, appropriately focused case studies could thus be of value (see pp. 78–9). As regards the RRA theory, a good deal of research has in fact been aimed at testing its consistency with results obtained from detailed interviews with samples of young people (e.g. Need and de Jong, 2000; Sullivan, 2006) or with their parents (e.g. Stocké, 2007) that focus on educational goals, plans and expectations. It has in this way been found *inter alia* that, while general attitudes towards education and

its intrinsic and extrinsic value differ little by class background, eventual educational choices are often, if not invariably, influenced by considerations of maintaining parental levels of both education and social class – as would be expected under the theory. At the same time, though, it has also been indicated through such research that there are other factors that may additionally serve to create secondary effects: for example, a tendency for students' assessments of their own ability to be higher the more advantaged their social backgrounds, even when previous performance is controlled, and also a tendency for informational as well as economic constraints to be greater for students from less advantaged backgrounds.

The second research strategy involves what could, in contrast, be described as attempts at the indirect observation of hypothesised causal mechanisms. In this case, the aim is to show that the mechanism under examination implies *other* regularities apart from those that it is intended to explain, and then to see if these regularities can be demonstrated.[5] With the RRA theory, a particularly good example of this strategy is provided by the work of Davies, Heinesen and Holm (2002) and Holm and Jaeger (2008). What these authors note is that under the RRA theory, the effect of parental background on children's educational choices should not be continuous throughout their educational careers but rather 'kinked', in that it should weaken once children have reached an educational level that gives them a high probability of avoiding downward mobility. Through analyses of data on students' transitions within the Danish educational system, it is then shown that these derived expectations from the RRA theory are to a large extent, even if not always, supported.

The third possible research strategy is experimental rather than observational. It may be that in the light of a proposed causal

[5] This strategy can be seen as entailing the 'hypothetico-deductive' method as classically proposed by Popper (1959). At the same time, though, it is dependent on other regularities being derivable from the theory under examination – which in turn lends force to what has become known as the 'Fisher dictum'. Cochran (1965) reports that when R. A. Fisher was asked how observational studies could best be made to yield causal conclusions, he replied, 'Make your theories elaborate': that is, potentially exposed to testing in as many different ways as possible.

mechanism, an experimental, or at least quasi-experimental, study can be designed, through which it can be assessed how far an intervention or 'treatment' (see the discussion of Chapter 8) has effects of the kind that would be expected if the mechanism were in fact in operation. That is to say, in this context an 'effects of causes' approach may appropriately be taken up (see Gelman and Imbens, 2013). As regards the RRA theory, no specific experimental test has so far been developed. However, a major study approximating an RCT in its design, and influenced in part by the RRA theory, is presently under way in Italy.[6] The effects are being investigated of providing students in a sample of secondary schools with specialist advice on their chances of success if they go on to university (given their academic performance to date), on the costs they are likely to incur in taking up particular courses and on the returns they are likely to gain. By then comparing the choices made by students who received this advice with those in a control sample of schools who received no advice, it will be possible to make some estimate of the importance of purely informational as distinct from economic constraints on the decision to enter higher education. Under the RRA theory, the expectation would be that, while some reduction in secondary effects in class inequalities may in this way be achieved, such effects will largely remain, since differences in the risks involved in this decision related to class inequalities in economic resources will still be in operation.

These different research strategies that may be followed in testing hypothesised causal mechanisms are not to be ranked in some order of importance. Each has its own advantages and disadvantages. What is important is that the actual operation of mechanisms should be tested *in as many ways as is possible* and the results obtained be considered in relation to one another.[7] It should not be expected that

[6] The study is directed by Professor Antonio Schizzerotto at the Research Institute for the Evaluation of Public Policy, Trento.

[7] One other possible strategy is that conducted via what has become known as agent-based computational (ABC) modelling. In this case (see Epstein, 2006: ch. 1), the basic idea would be to ask how a given population regularity could be generated through the actions and interactions of heterogeneous and autonomous agents, and then to

any particular test will produce 'clinching' results, at least not of a positive kind, but at best 'vouching' results – to take up Cartwright's (2007: ch. 3) useful distinction; and greatest weight has then to be given to how far results from different tests do or do not 'fit together'. In this regard, Haack's (1998: ch. 5) 'crossword-puzzle model' for the evaluation of evidence in relation to a hypothesis, emphasising the consistency or inconsistency of the implications of different empirical findings, would appear especially apt (see also Cox and Donnelly, 2011: chs 1, 2).[8]

The approach to mechanism-based explanations in sociology that starts out from some established population regularity as the explanandum does then tend to differ from the approach aiming to create a catalogue, or toolbox, of explanatory mechanisms in the importance that is attached to the question of whether a mechanism is that which is actually in operation in a particular instance – over and above the question of its causal adequacy. There is, moreover, one other difference that emerges between the two approaches that is of some consequence and that should in conclusion also be noted.

With the catalogue approach, a quite catholic view is taken – appropriately enough – as regards the theoretical basis of the mechanisms that are specified. Thus, Hedström and Bearman (2009a: 22, n. 1) point out that, although proponents of mechanism-based explanation in sociology do in general seek to specify mechanisms in terms

attempt to construct a model that could be shown, through computer simulation, to be capable of 'growing' the regularity in question. This strategy can provide a strong test of the causal adequacy – or of what ABC modellers refer to as the 'generative sufficiency' – of a proposed mechanism, and interesting and theoretically suggestive applications are now emerging in both sociology and demography (see e.g. Todd, Billari and Simão, 2005 for a model able to reproduce observed regularities in age at first marriage, based on 'fast and frugal' heuristics). However, to repeat the point made in the text, to show the generative sufficiency of a mechanism is not to show that it is in fact this mechanism that is in some particular instance at work.

[8] I am grateful to Jan Vandenbroucke for drawing my attention to Haack's work and also (together with David Cox) to a classic paper in epidemiology that provides an outstanding illustration of the crossword-puzzle model in application: the meta-analysis of the evidence for smoking as a cause of lung cancer by Cornfield et al. (1959).

of the action and interaction of individuals, this does not imply a similarly general commitment to rational-action theory. The mechanisms proposed as the 'nuts and bolts, cogs and wheels' of sociological explanations may be ones in which key importance attaches to action that is primarily orientated to the expectations of others and to conformity with the social norms that prevail within groups, social networks, communities and so on. However, insofar as sociologists concerned with population regularities of a well-established but still opaque kind have sought mechanism-based explanations of these regularities, the tendency has been for these mechanisms to be envisaged as entailing action that could indeed be understood as rational – albeit more often, as with the RRA theory, in a bounded rather than in a demonic sense.

The significance of this difference does then emerge in regard to what is perhaps the strongest objection that has thus far been put forward to the idea of mechanism-based explanation, both in general and in the social sciences in particular. This is the objection (see e.g. Kincaid, 2011; cf. King, Keohane and Verba, 1994: ch. 3) that seeking the generative mechanisms that underlie observed regularities leads in effect to an infinite regress. The philosopher Patrick Suppes observed some time ago that '. . . the mechanisms postulated and used by one generation are mechanisms that are to be explained and understood themselves in terms of more primitive mechanisms by the next generation' – or, in short, that 'one man's mechanism is another man's black box' (Suppes, 1970: 91). And the question can then be raised of whether in this process there is any evident stopping point – except perhaps through some appeal to 'covering laws' of nature (themselves inexplicable) of the kind that mechanism-based explanation is aimed at avoiding.

With regard to the social sciences, Hedström (2005: 27–8) has argued that appropriate stopping points can be identified: that is, where the mechanisms invoked are no longer ones that lie within the range of interest of these disciplines, but, presumably ones which extend into the biological or physical sciences. But a stronger response

is in fact possible. Insofar as mechanisms that are taken to explain population regularities appeal to action reflecting social norms, then, even if it can be shown that these mechanisms are indeed in operation, further questions remain open and need to be pursued: that is, as previously argued in Chapter 3, questions of why it is *these* norms rather than others that are influential and of why individuals do conform with these norms rather than contravening or perhaps openly challenging them. Until questions of this kind are answered, it could be held that black boxes clearly do exist. In contrast, insofar as the action involved in a mechanism can be treated as rational – even if only in a subjective, bounded sense of being seen by the individuals concerned as that best suited to attaining their ends, given the conditions under which they are required to act – a different situation obtains. In this case, a stopping point could be thought to have been reached in that, as Hollis (1977: 21; cf. Boudon, 2003a, Introduction) has put it, 'rational action is its own explanation'; or, as argued by Coleman (1986: 1), the rational action of individuals, even if the rationality is only subjective, is 'understandable' action that we need ask no more questions about and thus has 'a unique attractiveness' as the basis of sociological theory. In other words, if the 'bottom line' of a sociological explanation is not social norms but rather rational action – which may or may not result in conformity with norms – both explanatory *and* hermeneutic requirements are in this way met (see further Goldthorpe, 2007: vol. 1, ch. 7).[9]

[9] Watts has argued that explanations of social phenomena in terms of what he calls 'rationalizable action', while attractive in providing 'understandability', still 'cannot in general be expected to satisfy the standards of causal explanation' (Watts, 2014: 314–15). However, the standards he supposes are those of the potential outcomes approach, which, as maintained in the previous chapter, can be questioned at least as regards their applicability in sociology; and further, as his paper goes on, it appears to turn essentially into a plea for the testing of explanatory models that invoke 'rationalizable action' on an out-of-sample basis – that is, on the basis of data and analyses other than those which led to their initial formulation – which is of course entirely in line with the argument of this chapter. Further questions could of course be raised concerning the ends towards which rational action is directed. However, as observed in Chapter 2, how far individuals' choice of ends is open to systematic explanation of any kind remains a matter of serious doubt.

In the context of sociology as a population science, the search for mechanism-based explanations of established probabilistic regularities could then be said to proceed with two distinctive emphases. First, and consistently with a concern for the causes of effects, the emphasis is less on the effects that mechanisms *could* produce than on the testing of whether proposed mechanisms are those actually at work in particular cases of interest. Second, and consistently with the underlying individualistic paradigm, the emphasis is on mechanisms that can be ultimately expressed in terms of individuals' informed choices among the possibilities that they see as open to them and of the rationality involved in such choices and the action that follows from them.