Part I

Explanation and Mechanisms

This book relies on a specific view about explanation in the social sciences. Although not primarily a work of philosophy of social science, it draws upon and advocates certain methodological ideas about how to explain social phenomena. In the first three chapters, these ideas are set out explicitly. In the rest of the book they mostly form part of the implicit background, although from time to time, notably in the Conclusion, they return to the center of the stage.

I argue that all explanation is causal. To explain a phenomenon (an *explanandum*) is to cite an earlier phenomenon (the *explanans*) that caused it. When advocating causal explanation, I do not intend to exclude the possibility of intentional explanation of behavior. Intentions can serve as causes. A particular variety of intentional explanation is *rational-choice explanation*, which will be extensively discussed in later chapters. Many intentional explanations, however, rest on the assumption that agents are, in one way or another, *irrational*.[1] In itself, irrationality is just a negative or residual idea, everything that is not rational. For the idea to have any explanatory purchase, we need to appeal to specific forms of irrationality with specific implications for behavior. In Chapter 14, for instance, I enumerate and illustrate eleven mechanisms that can generate irrational behavior.

Sometimes, scientists explain phenomena by their *consequences* rather than by their causes. They might say, for instance, that blood feuds are explained by the fact that they keep populations down at sustainable levels. This might seem a metaphysical impossibility: how can the existence or occurrence of something at one point in time be explained by something that has not yet come into existence? As we shall see in Chapter 11, the problem can be restated so as to make explanation by consequences a meaningful concept. In the biological sciences, evolutionary explanation offers an example. In the social sciences,

[1] At this first occurrence in the book of the word "agent" it may be worthwhile to note that many scholars prefer "actor." Perhaps economists think in terms of agents, sociologists in terms of actors. Although it does not really matter which term we use, I prefer "agent" because it suggests agency; "actor," by contrast, suggests an audience that may or may not be present.

however, successful instances of such explanations are few and far between. The blood-feud example is definitely not one of them.

The natural sciences, especially physics and chemistry, offer *explanations by law*. Laws are general propositions that allow us to infer the truth of one statement at one time from the truth of another statement at some earlier time. Thus when we know the positions and the velocity of the planets at one time, the laws of planetary motion enable us to deduce and predict their positions at any later (or earlier) time. This kind of explanation is *deterministic*: given the antecedents, only one consequent (or antecedent) is possible. The social sciences offer few if any law-like explanations of this kind. The relation between explanans and explanandum is not one-one or many-one, but one-many or many-many. Many social scientists try to model this relation by using *statistical* methods. Statistical explanations are incomplete by themselves, however, since they ultimately have to rely on intuitions about plausible causal *mechanisms*.

1 Explanation

Explanation: general

The main task of the social sciences is to explain social phenomena. It is not the only task, but it is the most important one, to which others are subordinated or on which they depend. The basic type of explanandum is an *event*. To explain it is to give an account of why it happened, by citing an *earlier event* as its cause. Thus we may explain Ronald Reagan's victory in the 1980 presidential elections by Jimmy Carter's failed attempt to rescue the Americans held hostage in Iran.[1] Or we might explain the outbreak of World War II by citing any number of earlier events, from the Munich agreement to the signing of the Versailles Treaty. Even though in both cases the fine structure of the causal explanation will obviously be more complex, they do embody the basic *event-event* pattern of explanation. In a tradition originating with David Hume, it is often referred to as the "billiard-ball" model of causal explanation. One event, ball A hitting ball B, is the cause of – and thus explains – another event, namely, ball B's beginning to move.

Those who are familiar with the typical kind of explanation in the social sciences may not recognize this pattern, or not see it as privileged. In one way or another, social scientists tend to put more emphasis on *facts*, or states of affairs, than on events. The sentence "At 9 A.M. the road was slippery" states a fact. The sentence "At 9 A.M. the car went off the road" states an event. As this example suggests, one might offer a *fact-event* explanation to account for a car accident.[2] Conversely, one might propose an *event-fact* explanation to account for a given state of affairs, as when asserting that the attack on the World Trade Center in 2001 explains the pervasive state of fear of many Americans. Finally, standard social-science explanations often have a *fact-fact* pattern. To take an

[1] To anticipate a distinction discussed later, note that Carter did not *fail to attempt* but *attempted and failed*. A non-action such as a failure to attempt cannot have causal efficacy, except in the indirect sense that if others perceive or infer that the agent fails to act, they may take actions that they otherwise would not have or decide not to act when they otherwise would have acted.

[2] The voter turnout example discussed later provides another illustration.

example at random, it has been claimed that the level of education of women explains per capita income in the developing world.

Let us consider the explanation of one particular fact, that 65 percent of Americans favor, or say that they favor, the death penalty.[3] In principle, this issue can be restated in terms of events: How did these Americans *come to favor* the death penalty? What were the formative events – interactions with parents, peers, or teachers – that caused this attitude to emerge? In practice, social scientists are usually not interested in this question. Rather than trying to explain a brute statistic of this kind, they want to understand *changes* in attitudes over time or *differences* in attitudes across populations. The reason, perhaps, is that they do not think the brute fact very informative. If one asks whether 65 percent is much or little, the obvious retort is, "Compared to what?" Compared to the attitudes of Americans around 1990, when about 80 percent favored the death penalty, it is a low number. Compared to the attitudes in some European countries, it is a high number.

Longitudinal studies consider variations over time in the dependent variable. *Cross-sectional* studies consider variations across populations. In either case, the explanandum is transformed. Rather than trying to explain the phenomenon "in and of itself," we try to explain how it varies in time or space. The success of an explanation is measured, in part, by how much of the variation it can account for.[4] Complete success would explain all observed variation. In a cross-national study we might find, for instance, that the percentage of individuals favoring the death penalty was strictly proportional to the number of homicides per 100,000 inhabitants. Although this finding would provide *no* explanation of the absolute numbers, it would offer a *perfect* explanation of the difference among them.[5] In practice, of course, perfect success is never achieved, but the same point holds. Explanations of variation do not say anything about the explanandum "in and of itself."

An example may be taken from the study of voting behavior. As we shall see later (Chapter 14), it is not clear why voters bother to vote at all in national elections, when it is morally certain that a single vote will make no difference. Yet a substantial fraction of the electorate do turn out on voting day. Why do they bother? Instead of trying to solve this mystery, empirical social scientists usually address a different question: Why does turnout vary across elections? One hypothesis is that voters are less likely to turn out in inclement weather, because rain or cold makes it more attractive to stay home. If the data match

[3] Answers fluctuate. Also, the number of people who favor the death penalty for murder goes down drastically when life imprisonment without parole is stated as the alternative.
[4] Economists sometimes say that they are interested only in what happens "at the margin."
[5] Strictly speaking, the causal chain might go in the other direction, from attitudes to behavior, but in this case that hypothesis is implausible.

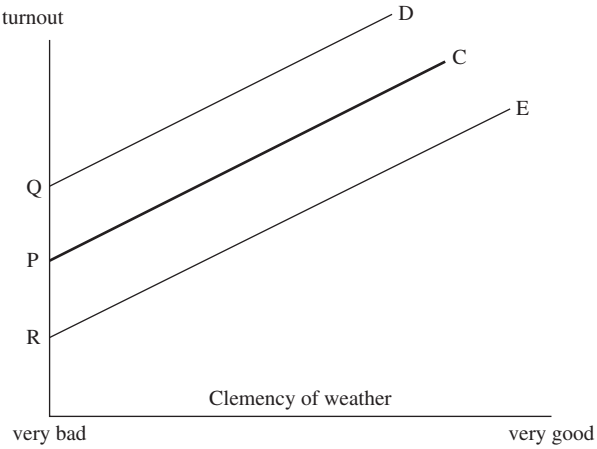

Figure 1.1

this hypothesis, as indicated by line C in Figure 1.1, one might claim to have explained (at least part of) the variation in turnout. Yet one would not have offered *any* explanation of why the line C intersects the vertical axis at P rather than at Q or R. It is as if one took the first decimal as given and focused on explaining the second. For predictive purposes, this might be all one needs. For explanatory purposes, it is unsatisfactory. The "brute event" that 45 percent or more of the electorate usually turn out to vote *is* an interesting one, which cries out for an explanation. I discuss it in several later chapters.

The ideal procedure, in an event-event perspective, would be the following. Consider two elections, A and B. For each of them, identify the events that cause a given percentage of voters to turn out. Once we have thus explained the turnout in election A and the turnout in election B, the explanation of the difference (if any) follows automatically, as a by-product. As a bonus, we might also be able to explain whether identical turnouts in A and B are accidental, that is, due to differences that exactly offset each other, or not. In practice, this procedure might be too demanding. The data or the available theories might not allow us to explain the phenomena "in and of themselves." We should be aware, however, that if we do resort to explanations of variation, we are engaging in a second-best explanatory practice.

Sometimes, social scientists try to explain *non-events*. Why do many people fail to claim social benefits they are entitled to? Why did nobody call the police in the Kitty Genovese case?[6] Considering the first question, the explanation

[6] The version of this episode that has entered the literature is the following. For more than half an hour on March 27, 1964, thirty-eight respectable, law-abiding citizens in Queens, New York,

might be that the individuals in question *decide* not to claim their benefits, because of fear of stigma or concerns with self-image. Since making a decision *is* an event, this would provide a fully satisfactory account. If it fails, social scientists would, once again, look at the *differences* between those who are entitled to benefits and claim them and those who are and do not. Suppose the only difference is that the latter are unaware of their entitlement. As an explanation, this is helpful but insufficient. To go beyond it, we would want to explain *why* some entitled individuals are unaware of their entitlement. To discover that because they are illiterate, they are unable to read the letters informing them about their rights would also be helpful but insufficient. At some point in the explanatory regress, we must either come to a positive event, such as a conscious decision not to become literate or a conscious decision by officials to withhold information, or turn to those who do seek the benefits to which they are entitled. Once we have explained the behavior of the latter, the explanation why others fail to seek their benefit will emerge as a by-product.

Considering the Kitty Genovese case, there is no variation in behavior to explain, since *nobody* called the police. Some accounts of the case indicate that several of the observers *decided* not to call the police. In terms of proximate causes this provides a fully satisfactory account, although we might want to know the reasons for their decision. Was it because they feared "getting involved" or because each observer assumed that someone else would call the police ("Too many shepherds make a poor guard")? Some of the observers, however, apparently did not even think about calling the police. One man and his wife watched the episode for its entertainment value, while another man said he was tired and went to bed. To explain why they did not react more strongly one might cite their shallow emotions, but that, too, would be to account for a negative explanandum by citing a negative explanans. Once again, their behavior can only be explained as a by-product or residual. If we have a satisfactory explanation of why some individuals thought about calling the police, even if in the end they decided not to, we shall have the only explanation we are likely to get of why some did not even think about it.

In the rest of this book I shall often relax this purist or rigorist approach of what counts as a relevant explanandum and an appropriate explanation. The

watched a killer stalk and stab a woman in three separate attacks in Kew Gardens. Twice their chatter and the sudden glow of their bedroom lights interrupted him and frightened him off. Each time he returned, sought her out, and stabbed her again. Not one person telephoned the police during the assault; one witness called after the woman was dead. Although recent research has shown that the version is factually incorrect, the general phenomenon of bystander passivity is well documented (Chapter 12). In references to the case in later chapters I assume the erroneous version, which has become part of the folklore of scholarship. I shall put "Kitty Genovese" in quotation marks, however, to remind the reader that it is a proxy for a more general and better documented class of phenomena.

insistence on event-focused explanations is a bit like the principle of methodological individualism, which is another premise of the book. In principle, explanations in the social sciences should refer only to individuals and their actions. In practice, social scientists nevertheless refer to supraindividual entities such as households, firms, or nations, either as a *harmless shorthand* or as a *second-best approach* forced upon them by lack of data or of fine-grained theories.[7] These two justifications also apply to the use of facts as explananda or as explanantia, to explanations of variation rather than of the phenomena "in and of themselves," and to the analysis of negative explananda (non-events or non-facts). The purpose of the preceding discussion is not to hold social scientists to pointless or impossible standards, but to argue that at the level of first principles the event-based approach is intrinsically superior. If scholars keep that fact in mind they may, at least sometimes, come up with better and more fruitful explanations. When we try to explain the decisions made at the Federal Convention of 1787, the recorded votes of the state *delegations* are useful, but incomplete. Historians have improved our understanding by identifying the votes cast by individual *members* of these delegations. Explanations of why the German National Assembly in 1933 and the French National Assembly in 1940 abdicated their powers gain much in power and focus when we can trace the changing and interacting motivations of individual deputies.

Sometimes, methodological individualism should force us to lower our sights. Social scientists are naturally drawn to big questions, yet some questions may be too big to allow for an answer. We may be able to explain the rise of Calvinism, but not the existence of some form of religion in virtually all societies. We may be able to explain the emergence of capitalist forms of agriculture in eighteenth-century England, but not the "transition from feudalism to capitalism" in Europe as a whole. Discussions of "the Axial age" and "modernity" also flounder, among other reasons, for lack of identifiable agents and their motivations. If social scientists are enjoined to use the microscope rather than the telescope, some questions may of course elude them forever. The loss in breadth is offset, or more than offset, by the gain in depth.

[7] Two economists, correctly observing that "neo-classical utility theory applies to individuals and not to households," set out to explain consumer behavior by appealing only to the preferences of individuals instead of the traditional household-centered approach. Nevertheless, they assume that family decisions are Pareto-efficient, implying that bargaining never breaks down. In real households, however, wives and husbands or parents and children often fail to reach Pareto-efficient decisions, because they do not agree on the division of the jointly created surplus. I mention this not as an objection to their work, which does indeed go beyond the traditional models, but to show that it can be difficult to apply methodological individualism in the absolutely literal sense.

Sometimes, we might want to explain an event (or rather a pattern of events) by its consequences rather than by its causes. I do not have in mind explanation by *intended* consequences, since intentions exist prior to the choices or actions they explain. Rather, the idea is that events may be explained by their *actual* consequences, typically, their *beneficial* consequences for someone or something. As a cause must precede its effect, this idea might seem to be incompatible with causal explanation. Yet causal explanation can also take the form of explanation by consequences, if there is a loop from the consequences back to their causes. A child may initially cry simply because it feels pain, but if the crying also gets it attention from the parents, it may start crying more than it would have done otherwise. I argue in Chapter 11 that this kind of explanation is somewhat marginal in the study of human behavior. In most of the book, I shall be concerned with the simple variety of causal explanation in which the explanans – which might include beliefs and intentions oriented toward the future – precedes the occurrence of the explanandum.[8]

In addition to the fully respectable form of functional explanation that rests on specific feedback mechanisms, there are more disreputable forms that simply point to the production of consequences that are beneficial in some respect and then without further argument assume that these suffice to explain the behavior that causes them. When the explanandum is a *token*, such as a single action or event, this kind of explanation fails for purely metaphysical reasons. To take an example from biology, we cannot explain the occurrence of a neutral or harmful mutation by observing that it was a necessary condition for a further, advantageous one. In a rare moment of methodological sobriety, Marx refers to the speculative distortions by which "later history is made the goal of earlier history, e.g. the goal ascribed to the discovery of America is to further the eruption of the French Revolution." In a less sober moment, he wrote that "The anatomy of man is the key to the anatomy of the ape."

When the explanandum is a *type*, such as a recurrent pattern of behavior, it may or may not be valid. Yet as long as it is not supported by a specific feedback mechanism, we should treat it as if it were invalid. Anthropologists have argued, for instance, that revenge behavior has beneficial consequences of various kinds, ranging from population control to decentralized norm enforcement (Chapter 21 offers many other examples). Assuming that these benefits are in fact produced, they might still obtain by accident. To show that they arise non-accidentally, that is, that they sustain the revenge behavior that causes them, the demonstration of a feedback mechanism is indispensable.

[8] For some purposes, it may be useful to distinguish among causal, intentional, and functional explanation. Physics employs only causal explanation; biology additionally admits functional explanation; and the social sciences further admit intentional explanation. At the most fundamental level, though, all explanation is causal.

And even when one is provided, the initial occurrence of the explanandum must be due to something else.

The structure of explanations

Let me now turn to a more detailed account of explanation in the social sciences (and, to some extent, more generally). The first step is easily overlooked: before we try to explain a fact or an event we have to establish that the fact *is* a fact or that the event actually did take place. As Montaigne wrote, "I realize that if you ask people to account for 'facts,' they usually spend more time finding reasons for them than finding out whether they are true . . . They skip over the facts but carefully deduce inferences. They normally begin thus: 'How does this come about?' But does it do so? That is what they ought to be asking."

Thus before trying to explain, say, why there are more suicides in one country than in another, we have to make sure that the latter does not tend, perhaps for religious reasons, to underreport suicides. Before we try to explain why Spain has a higher unemployment rate than France, we have to make sure that the reported differences are not due to different definitions of unemployment or to the presence of a large underground economy in Spain. If we want to explain why youth unemployment is higher in France than in the United Kingdom, we need to decide whether the explanandum is the rate of unemployment among young people who are actively searching for jobs or the rate among young people overall, including students. If we compare unemployment in Europe and the United States, we have to decide whether the explanandum is the unemployed in the literal sense, which includes the incarcerated population, or in the technical sense, which only includes those searching for work.[9] Before we try to explain why revenge takes the form of "tit for tat" (I or one of mine kill you or one of yours each time you or yours kill one of mine), we should verify that this is actually what we observe rather than, say, "two tits for a tat" (I kill two of yours each time you or yours kill one of mine). Much of science, including social science, tries to explain things we all know, but science can also make a contribution by establishing that some of the things we think we know simply are not so. In that case, social science may also try to explain *why* we think we know things that are not so, adding as it were a piece of knowledge to replace the one that has been taken away.[10]

[9] In either of the last two cases, some individuals may take up a career as criminals or students because they do not think they would get a job if they tried. For some purposes, one might want to count these among the unemployed; for other purposes, not.

[10] Just as science can help explain popular beliefs in non-facts, it can help explain popular beliefs in false explanations. For instance, most of those who suffer from arthritis believe arthritic pain

Suppose now that we have a well-established explanandum for which there is no well-established explanation – a *puzzle*. The puzzle may be a surprising or counterintuitive fact, or simply an unexplained correlation. One small-scale example is "Why are more theology books stolen from Oxford libraries than books on other subjects?" Another small-scale example, which I shall explore in more detail shortly, is "Why do more Broadway shows receive standing ovations today than twenty years ago?"

Ideally, explanatory puzzles should be addressed in the five-step sequence spelled out in the following. In practice, however, steps (1), (2), and (3) often occur in a different order. We may play around with different hypotheses until one of them emerges as the most promising, and then look around for a theory that would justify it. If steps (4) and (5) are carried out properly, we may still have a high level of confidence in the preferred hypothesis. Yet for reasons I discuss in the next chapter, scholars might want to limit their freedom to pick and choose among hypotheses.

1. Choose the theory – a set of interrelated causal propositions – that holds out the greatest promise of a successful explanation.
2. Specify a hypothesis that applies the theory to the puzzle, in the sense that the explanandum follows logically from the hypothesis.
3. Identify or imagine plausible accounts that might provide alternative explanations, also in the sense that the explanandum follows logically from each of them.
4. For each of these rival accounts, refute it by pointing to additional testable implications that are in fact *not* observed.
5. Strengthen the proposed hypothesis by showing that it has additional testable implications, preferably of "novel facts," that are in fact observed.

These procedures define the *hypothetico-deductive method*. In a given case, they might take the form shown in Figure 1.2. I shall illustrate it by the puzzle of increasing frequency of standing ovations on Broadway. It is not based on systematic observations or controlled experiments, but on my casual impressions confirmed by newspaper reports. For the present purposes, however, the shaky status of the explanandum does not matter. If there are in fact more standing ovations on Broadway than there were twenty years ago, how could we go about explaining it?

is triggered by bad weather. Studies suggest, however, that there is no such connection. Perhaps we should drop the search for the causal link between bad weather and arthritic pain and instead try to explain why arthritics believe there is one. Most likely they were once told there was a connection and subsequently paid more attention to instances that confirmed the belief than to those that did not.