Problems and Methods in the Study of Politics

Edited by

Ian Shapiro, Rogers M. Smith, and Tarek E. Masoud

External & Internal Explanation

CAMBRIDGE
UNIVERSITY PRESS

2004

7 External and internal explanation

John Ferejohn

I. Introduction

Should the social sciences focus more than they now do on solving real (explanatory) problems and less on developing methodologies or pursuing methodological programs? Two distinct worries animate this question. One is that too many resources may be devoted to the development and refinement of methodologies and theories, while too little attention is paid to the actual things needing explanation. In this sense there may be a misallocation of social scientific resources. The other worry is that when proponents of some methodology turn to explaining a particular event or phenomenon, they tend to produce distorted accounts; they are deflected by their inordinate attention to and sympathy for their favorite method. Method-driven social science comes up with defective explanations. Proper attempts to explain things, one might think, ought to be open ended and responsive to the phenomenon to be explained and not be committed in advance to any particular explanatory methodology. Such a commitment smacks of dogmatism or *a priori*-ism. These complaints are often illustrated by the familiar metaphors of drunks searching under street-lamps and the law of the hammer.

My inclination is to resist the question as not quite usefully posed. The development of systematic methodologies and theories is what permits the social sciences – or particular approaches to social science – to make distinctive and sometimes valuable contributions to understanding the events that interest us.[1] There are several reasons why this is the case. A methodological focus can throw new light on old issues in various ways; things that might be taken for granted from one perspective look problematic and in need of explanation from another. It can show how new kinds of evidence can bear on the explanation of an event, and how evidence – old as well as new – ought be interpreted. Even if commitment to a particular method tends to produce uneven or partial explanations

[1] For purposes of this chapter, I shall use "theory" and "methodology" interchangeably.

in some cases, such a commitment can enhance our understanding of phenomena by providing a new perspective on events that had previously been thought to be adequately understood.

Indeed, one might ask whether there is any sense to the idea of "an" explanation of any particular phenomenon. If one accepts the notion that explanations are answers to questions and that questions themselves are dependent on methodology, the notion of a unique or best or privileged explanation seems hard to defend. The best we can hope for is to pose method-driven questions about an event of interest (for whatever reason) and then try to make explanatory headway from the viewpoint of that question. By examining an event from more methodological perspectives, we can hope to increase our "understanding" or grasp of it. This is not to deny progress in understanding but it is to say that understanding is a multifaceted idea and that improving the understanding of an event does not imply the existence either of a unique explanatory viewpoint (or methodology) or of a best explanation.[2] In this sense, we cannot avoid methodology any more than Molière's gentleman could avoid speaking prose.

So, as a general matter, methodological eclecticism seems not only defensible, but unavoidable. But perhaps we cannot approach all interesting or important events eclectically. It is often thought that certain kinds of historically important events – "singular" events – are inaccessible to method-driven social science. To understand such occurrences we must resort to other forms of understanding, especially techniques of narrative or interpretive reconstruction. There may be some truth to this claim if we identify methods-driven social science with the pursuit of empirical regularity and causal explanation, of the kind sought by practitioners of the natural sciences. But I think that is much too narrow a conception of social science. Of course we are interested in causal explanation but also

[2] In fact, methodological programs generate events needing explanation. This idea is a commonplace among those social scientists who generate their data in a theory-based manner – as did, for example the sociologists who defined and operationalized the concept of alienation or the economists who formulated the idea of consumer confidence with survey research methodology. While both of these ideas may have originated as explanatory variables to account for other phenomena, they have long since become free-standing dependent variables needing explanation.

Besides, what is an "event" anyhow? We often understand an event requiring explanation as a kind of primitive pretheoretic idea but that is, at best, just a rough starting point. Event descriptions are themselves dependent on method or theory. Events are theoretically unstable in that as accepted theory changes, the events needing explaining change too, as the voting example suggests. And, if any doubt remains one can point to the ubiquitous use of prisoners' dilemmas and collective action problems to turn Aristotle's axiomatic conception of man as inherently social on its head. Group formation and social organization become things in need of explanation rather than givens.

in other questions as well. Perhaps considering the example of singular events will help sharpen some of these other explanatory aims, and show that methods-driven social science can help in their pursuit.

There are so many different methodological programs within the social sciences that it would be impossible and distracting to try to discuss even a fraction of them. But much of what I have to say can be illustrated by considering two broad approaches to explanation found within the social sciences. Externalists explain action by pointing to its causes; internalists explain action by showing it as justified or best from an agent's perspective. Externalist explanations are positivist and predictive; internalist explanations are normative or hermeneutic. Externalists tend to call themselves political scientists; internalists, political theorists. And, both externalists and internalists agree, if they agree on little else, that they are engaged in different enterprises.

I shall argue against this widely shared belief. I believe that what distinguishes social science from humanistic accounts of action is a recognition of the importance and validity of externalist explanation – seeking to understand events by finding causal information about them. And, what distinguishes social and non-social scientific explanatory practice is a recognition of the importance and validity of internalist accounts: seeing actions as more or less justified from the perspective of normatively guided actors. I do not think either of these commitments can be surrendered without giving up much of the distinctiveness and vitality of the social sciences. This is not to deny that there are differences between these broad methodological approaches, nor that these differences can run deep. But, it is to insist that the two approaches have in common a shared concern to explain roughly the same range of social phenomena. And I think that this shared concern often produces significant pressures for convergence in practice if not in theory.

II. Singular events: an example

Singular events are sometimes understood to mark the beginnings or ends of eras, to punctuate longer periods of slow or evolutionary development, or to carry special meaning or significance as events. Certain especially important wars, rebellions, assassinations, historic compromises, or elections supply typical examples. Within American political history, for example, the elections of 1800 (sometimes called the "revolution" of 1800), 1828, 1896, or 1932 might each qualify on grounds of their significance at the time or later. Truman's decision to use an atomic weapon against the Japanese, or his Marshall plan would probably count too, at least from some perspectives. And, of course, the outbreak of

certain wars, the American Civil War especially but also the American Revolution, are often thought of as singular and calling for a special kind of explanatory strategy.

Singular events are distinctively heterogeneous – each is supposed to be singular after all – and so one is entitled to doubt that much can be said about them as such as there may be nothing, other than their rarity and their *ex-post* significance, they have in common. We may well be forced to think of such events as bearing no more than a relationship of family resemblance to one another. Moreover, because such events are thought to be freighted with meaning – indeed having some kind of special meaning is part of why the event is given this status – the designation of an event as "singular" is bound to be controversial and "theory dependent."

If this is so, singularity is not really definable in an extensional sense. Events are typically taken to be singular partly because of what they mean or signify, or because of some consequences that follow from them. The concessions to Hitler at Munich in 1938 are thought to be singular, for example, partly because of Hitler's subsequent invasions of Poland and Western Europe, and partly because of the horrific character of his regime. Had he been run over by a car late in 1938, or had he turned out to be a small-time bully who had no further territorial ambitions, Munich itself would probably have faded from view and lost any claim to special importance. Munich came to be characterized as an act of appeasement of a tyrant – indeed it has come to stand for any such act – because of how Hitler acted afterwards. The significance of Munich then rests on the beliefs that Hitler had an insatiable lust for domination, that he was there given extra time and opportunity to develop his forces for future aggressions, and that the results of these aggressions turned out to be momentous for the rest of Europe. In this respect, the singularity of Munich depends both on subsequent events and shared meanings.

The explanation of such occurrences seems especially difficult from a scientific viewpoint for several reasons. The first issue is, as already mentioned, that such events however identified have nothing in common.[3] I am happy to grant this because not much turns on it; there is no reason to think that good explanations of singular events would necessarily have much in common either. The designation "singular" only conveys rarity and special significance for a people. Nothing more than that. But the most important practical explanatory difficulty seems to be that singularity itself implies that there is a relatively thin base of information to draw

[3] This is so in two senses: first, their meanings and subsequent consequences are heterogeneous. And second, they are identified by reference to shared beliefs or meanings and this implies that singularity is not robust to changing beliefs or understandings about the event.

on – at least thin in some important sense. In another sense, we might have quite a lot of narrative information about such an event precisely because of its importance to contemporaries or successors. Participants will tend to save their papers, write memoirs or produce apologies; historians and others will try to understand and explain it; politicians will invoke it as justifying some further course of action.

So, we will tend to have at once both too little and too much information about a singular event. Moreover, our capacity to explain or understand such events sometimes is important from a policy-oriented standpoint. Witness, for example, the role of "Munich" in debates about how the United States or the United Nations should deal with Saddam Hussein. Those who cite Munich argue that a particular "causal" process was in motion in contemporary Iraq – roughly the same kind of process that was in motion in Germany in 1938 – that would have led to the development or acquisition of weapons of mass destruction unless there was an outside intervention to put a stop to it. This argument rested on the claim that we have an explanation or understanding of developments in Germany in the late 1930s that instructs us as to how to behave toward (relevantly similar) new tyrants. But how could this understanding be relevant to Iraq, given the assumed singularity of Munich?

One strategy of explanation is to embed Munich (and Saddamite Iraq) in a class of relevantly similar cases and then attempt to find generalizations that seek to connect, in some standard sense, appeasement and its consequences. This is, in effect, to deny that Munich was actually a singular event – or at any rate to make its singularity explanatorily invisible. Large-N studies of the outbreak of wars and rebellions – many of which might count as singular when viewed from another perspective – are examples of this strategy. Another approach is to break up the larger event into complex sequences and ensembles of actions – subevents – where these sequences and subevents are more mundane and perhaps explicable by resorting to ordinary explanatory strategies.[4] For example, in explaining the American response to the placement of Russian missiles in Cuba, Graham Allison (1971) draws heavily on the organization theories of the time. From these more or less ordinary behavioral predictions

[4] There are a lot of ways this might be done. One is to break the event into its sequential happenings, horizontally across time. Another is to layer the event vertically into simpler subprocesses. One lesson of Allison's (1971) analyses of the Cuban missile crisis is that there are many ways in which this may be done. Another issue is what kinds of external background information, theoretical or empirical, may be usefully called upon to analyze the patterns of subevents. Allison himself draws upon a complex mixture of theoretical knowledge from various areas of social science in order to cast light on the complex events associated with the crisis. The fertility of this approach is perhaps best seen from the illuminating responses of critics.

he crafts or aggregates a story of the crisis and its resolution. This strategy involves assuming that the singularity of the event does not infect its subevents in any strong way. Allison assumed that more or less ordinary regularities of organizational behavior would be exhibited throughout the missile crisis. It was in the concatenation of these events that the singularity of the crisis arose.

Both of these explanatory strategies seem completely compatible with ordinary social science practice and I have nothing to say against their pursuit. There is nothing in the nature of the explanations offered that makes them different, in principle, from any other social scientific explanation. The thinness of the data and plausibility of the comparisons, and indeed the cleverness of the explanatory strategies, may tend to make the particular accounts incompletely convincing and controversial but the resulting disagreements are not in any way mysterious. But both approaches involve, in one way or another, denying the assumption that the event in question is in some way singular.

There is another familiar strategy of explanation of singular events that seems more unusual from the standpoint of ordinary social science explanation and which is predicated directly on singularity itself. This strategy does not depend on constructing larger classes of similar events, or on breaking down the event into smaller sequences. It focuses instead on explaining the event from the point of view of the actor or actors involved. The availability of this approach, everyone will quickly notice, does not necessarily turn on the singularity of the event at all but rather on our access to the perspectives of involved agents. Any action, however ordinary, could be examined from the viewpoint of the agent or agents involved if we had some way of getting access to those viewpoints. In the case of singular events, however, it seems to me that this strategy – of agent-centered explanation – seems to recommend itself especially strongly to historians and others who are convinced of the event's significance. Who better to consult about Munich than Chamberlain or Hitler or perhaps those of their agents who were present for the discussions? Those involved seemed have a special access to what was said and thought and, perhaps, to whatever it is that made Munich singular (though this last claim seems disputable if singularity arose partially from later events that could not be foreseen).

Moreover, when an event is seen to be of special importance, participants and close observers are probably more likely to remember what they did, to save their records, to speculate and explain, justify, and criticize what happened. Such recollections will need to be treated with caution because participants may have scores to settle, people to protect, and further agendas to pursue. Or it may be that memory processes are

particularly likely to distort recollections in particular and invidious ways. Participants may, like the rest of us, have a tendency to reconstruct events that they think they are remembering. Still, such information, if sensitively handled, might seem especially privileged material for explanatory purposes.

Narrative access facilitates the resort to a deliberative perspective on the event. From an agent's perspective it is often quite natural to invoke some idea of what would have happened (or what the agent believed or should have believed would have happened) in counterfactual circumstances to explain (or justify) her choices. Of course, we can have little evidence about the accuracy of counterfactual conjectures. But from an agent-centered viewpoint, that does not matter very much. To explain her actions, we need to invoke what she thought would have followed from alternative courses of action. Her beliefs can be criticized of course, and if we think that they were badly formed or incoherent, that might affect our assessment of the overall course of events. One might think about Munich specifically, that there was, in fact, a lot of motivated belief formation occurring – that participants came to believe what they wanted to believe – and that such phenomena are characteristic of events like Munich. Of course, insofar as beliefs at the time evolved under these pressures, we might think that those beliefs came under quite distinct pressures as subsequent events unfolded.

In any case, the worry is that narrative or agent-centered explanation might be overly seductive and therefore tend to crowd out other kinds of explanations. This is not worrying if there is reason to think that each of these explanatory strategies will converge in some sense. But I think there are some reasons to doubt such a convergence and to think that agent-centered explanations will tend to diverge systematically from the comparative and counterfactual approaches outlined above. Are these reasons good? Or do they turn on misapprehensions about either social scientific or agent-centered explanation?

III. Two approaches to explanation

I will distinguish two kinds of explanation – external or more or less causal explanation, and internal or deliberative explanation. External explanations represent agents as doing things because of some configuration of causal influences, broadly speaking. I mean to include not only ordinary causal explanations but also functional or structural explanations as well. Functional or structural accounts imply the existence of some vindicating causal processes; functional explanations in biology, if true, are vindicated by the existence of causal processes described under the rubric of natural

selection. Structural explanations, to be complete, require some underlying causal processes standing ready to police structural irregularities. So, for present purposes I shall speak only of causal explanation where it is to be understood that I use this expression to stand for a broader class of social scientific accounts.

Obviously, the notion of "cause" needs to be spelled out and how it is spelled out will turn out to depend on the theory of behavior that is supposed to be at work in the agents. But, whichever theory is in play could work in one of two general ways: one, causal factors might determine behavior without "running through" the agent's deliberative or reasoning processes. When a doctor hits your knee with a hammer, it moves automatically. Or, causal factors may drive or shape the agent's deliberations by presenting her with reasons for action so that she "chooses" certain actions rather than others based on those reasons (this is developed in more detail in Pettit (1993)). Either way, we could understand her action as caused, but in the latter case she is also acting deliberatively (for reasons) as well.

We may consider some more or less classical examples of external explanations: Durkheim's account of societal suicide rates. Some fraction of people in certain kinds of societies will, Durkheim thought, tend to kill themselves at certain rates and these rates are said to be predictably related to certain attributes of the society. So, one might want to say that a society's suicide rate is explained causally. It is not clear which kind of causal explanation this is supposed to be, but it seems that it could be causal in the second (deliberative) sense. In any case, for this to be a causal explanation of any kind seems to require that there are mechanisms that tend to lead the particular individuals who commit suicide to do this. It is possible that a mechanism that produces this result might work by presenting individuals with reasons and that the individuals themselves all act deliberatively.[5] That the rate turns out to be predictable, while the individual event is not, might reflect the fact (if it is a fact) that the aggregate-level causal factors relate statistically to the circumstances of the individuals.

Drug addiction offers a different kind of example. Here, the agent does something compulsively, in spite of her judgment as to what she thinks she should do or in spite of an intention not to do it. There are various ways to understand this phenomenon. There may be deliberation and choice involved even though the outcome of the deliberative process is a

[5] Persons in certain social locations, for example, might find themselves faced with particular conflicts among reasons and there may be more such locations in some societies than in others.

predictable consequence of the working of a chemical addiction. The drug might work, for example, by altering the preferences of the individual so that she has a strong reason to take the drug when in certain chemical induced circumstances. Or, the drug might work by suspending some or all of the agent's deliberative capacities so that she takes the drug in spite of knowing that she has sufficient reason not to do so. In the first case, the drug works through agent deliberation; in the second, it works around agent deliberation. But either way, there is a causal relation between having the addiction and consuming the drug.

Internal or deliberative explanations work differently. An internal explanation for an action identifies reasons that an actor had that would rationally lead to or produce the action. An action is explained internally as an outcome of a deliberative process in which the agent is assumed to act for reasons; to take actions which are best on the reasons available to the agent. To "explain" in this sense is to "justify," in that an actor with the goals and beliefs held by (or attributed to) the actor would have, or could rationally have adopted that action. Internal explanation is, in this respect, a normative account of an agent's behavior in that an action is explained only if there are sufficient reasons, from the agent's viewpoint, for doing it. From the agent's perspective, the action was the best thing to do.

There are some qualifications and ambiguities that need to be addressed from the internal perspective that may be important later on. One is that an actor can have reason to do X, and do X, but do it for some other reason. I might have reason to go to Chicago tomorrow to complete some transaction, but I might actually go to Chicago because I was abducted by kidnappers at gunpoint. In this case I have two sufficient reasons to do X and so, in either case, X is justified and therefore internally explained. But there are, however, two different explanations and they are not compatible. In this case, I suppose it is clear that the reasons emanating from the kidnapper "trump" the other reasons and so the kidnapper-centered explanation is the correct internal account.[6]

Internal explanations do not involve any comparative test, at least not directly or obviously. We explain an agent's behavior by considering

[6] Some will want to say that the kidnapper-centered explanation, in which I comply with the kidnappers' requests because they are pointing guns at me, is causal. I am with Hobbes on this issue: the agent might have chosen to die or to risk death even if these were unattractive choices. Going along with kidnappers is a choice which, no doubt, has an awful lot to recommend it, and the agent's complying with the kidnappers' request is explained by its being the best course of action on the reasons presented. So, I resist the idea that this is a causal explanation. If one takes sufficiently compelling internal accounts to be causal, it is difficult to separate this example from one in which the kidnappers bop the person on the head, stuff them in a bag and dump them in Chicago, which does strike me as a causal story.

whether she had sufficient reason to take the action and, in fact, took the action because of that reason. This may involve appeal to the agent's beliefs about counterfactual circumstances, of course, and perhaps this would include her beliefs about what would happen in those circumstances. She may decide to act in order to bring about some desired state of affairs or to prevent the occurrence of some event. Still, the explanation itself does not appeal to anything other than the agent's own reasons for action, and not to what other agents would do in similar circumstances.

Examples: When we seek to understand or explain everyday events – why you were late to dinner last night – we normally seek an internal explanation. You say that your daughter needed help with her homework and you needed to delay your departure from your house for a few minutes in order to assist her. Your belief that your daughter was in need of assistance (together with your belief in the unimportance of being a bit late for dinner) provides a reason for delay and is the reason you acted on in leaving your house as late as you did. This account provides a justification for your leaving your house late, and an explanation to me for your tardiness. Obviously, the explanation in this case is also meant to excuse your lateness too. To say that we expect an internal explanation in this case does not mean that external accounts are unavailable. Presumably we would be reluctant to accept as true an internal explanation unless we believed that agents like you would generally find it reasonable to respond to a request for homework help. But it would be odd for you to offer the external account as an account of your action.

Ethnographic explanations are typically internal. For example, Richard Fenno (1978) explains how each of his Congressmen tried to present themselves, in their districts, by appealing to reasons they gave or appeared (to the analyst) to have and to have acted on. More elaborate ethnography seeks to explain the system of reasons that are available for motivating action in a community or culture. In his book, Fenno hoped also to provide an external account: he argued that one might be able to explain which kinds of Congressmen and/or which kind of congressional districts would tend to produce which kind of "home style." If he is right, the internal explanations that he offered might converge on a unified external or causal account of homestyles.

But there are reasons to be skeptical of such a convergence, at least when put forward as a general claim. The main reason is that there seems to be a fundamental distinction between causal accounts and justificatory accounts of action. Causal accounts of how things go seem to be, in their nature, positivistic in that they are offered as true accounts of the phenomena to be explained. Internal explanations or justifications seem to be inherently normative and so not to be anchored in the way the world

is. They certainly don't offer causal explanations for the action and would not be expected if a causal explanation was sufficient. If you could not help doing X, there is no need to give reasons or justifications for doing X. This is rough worry, and one that I will argue against, but I think it captures part of the skepticism about convergence. Moreover, there are a lot of people – social scientists, lawyers, philosophers – who seem to think that one or another of these two kinds of explanations is either impossible or unimportant. One area where this battle has raged is in law.

For example, consider the problem of explaining the structure of contract law within the United States. There are roughly two approaches to this problem within the legal academy: one account, law and economics, claims that most of the structure of this law can be explained as the structure of legal rules that would produce the most total wealth in the society. The law and economics of contracts explains how contracts are made, which contracts can be breached, what damages should be paid, which kinds of evidence count in ascertaining the existence of a contract and so on. The alternative approach explains these doctrinal elements as the consequences of shared and deeply held values that are at stake in contracting. An explanation for a rule, in this latter case, is a justification of the rule in terms of values embodied in the legal system. Unsurprisingly, each approach is better at explaining some legal structures than others and, for this reason, proponents of each approach sometimes try to fuzz the distinctions between them. Internalists sometimes recognize that wealth maximization may itself be a value and externalists are sometimes forced to endorse broader sets of values than wealth maximization as the driving causal motor for their theories. But the important point here is that proponents of each side tend to think of arguments from the other viewpoint as irrelevant, irreverent, and unimportant.

Part of the reason for this disagreement lies in different attitudes about causation in human affairs. Internalists, at least since Aristotle, see actions as taking place for reasons and tend to admit the possibility that an agent may fail to do what she has reason to do. Indeed, the concerns of internalists are at least partly driven by normative or moral concerns and a causal account of human action tends to undermine the kind of human agency needed to make normative appraisal relevant. Externalists are concerned to explain regularities they see in the social world and tend to believe that the only or the best way to account for such regularities is by appealing to more or less general causal mechanisms. From an externalist viewpoint, normative concerns are wholly separate from explanatory ones.

Some have argued in favor of a practical preference for seeking external rather than internal explanations when both kinds of explanation might be available. Gary Becker (1976), for example, believes that actions

are in fact brought about by agents seeking preference satisfaction given their beliefs in circumstances of constraint or scarcity and, in this sense, would be expected to believe that deliberative or internal accounts of economic activity are freely available. But he also thinks that because 'we' (economists) cannot really know much about the preferences or beliefs held by individual agents – we have no special access to these mental attitudes – good economic explanations will focus on showing how it is that the constraints do most of the causal work in producing actions, and making little appeal to subjective or deliberative factors. As far as any particular agent is concerned, the constraints themselves are likely be produced causally (through variations in prices or technologies). Thus, as I understand him, Becker has a practical preference for external or causal explanation even when deliberative or reason-based explanations are, in principle, available.

IV. Are internal and external accounts necessarily in conflict?

Superficially, external explanations claim to be more or less causal accounts whereas internal explanations explain actions by making them intelligible – showing how they are justified from the agent's perspective – which doesn't seem to be a causal story. One might try to convert internal accounts into causal accounts by taking reasons as causes. This would involve saying that if an agent has most reason to do X in circumstance C, that X is caused by facts about the agent and her circumstances. On the face of it this seems to involve denying that the agent has a capacity to choose not to do X. Or, it is to deny agency. Is this objectionable?

Maybe not. We are, after all physical/biological creatures embedded in the material world and, in that sense, whatever it is that our bodies do must be compatible with physical laws. So, if an agent does X (where X is described physically), there seems to be little objection to saying that X is caused. But, there are reasons to think that X, taken now as an action, cannot be completely described physically. The description of an act also includes the intentions with which it is done – waving my hand (a bit of behavior) might be a signal (an act) if I have a certain intent, or it might be part of an effort to dry my hands (a different act) if that is what I intend, or it might simply be a twitch (a mere bit of behavior) – the physical description doesn't tell us whether X is an act or what act it is. So, while there is no problem with saying that the physical aspect of X is caused, it is a stronger claim to say that X as an act is caused. Whether that stronger claim is sustainable depends on whether we are willing to attribute causal force to intentions.

There two issues here: first, what kind of linkage is there between intentions and actions? One might be willing to say that once an intention to do X is formed, the act of doing X follows automatically. This might make sense for certain kinds of simple intentions and actions – intending to pick up a fork might be so closely connected to picking it up that there is little room for deliberation or choosing between the intention and the action. But most acts of interest to us as social scientists are much more complex. If I form an intention to vote on election day, for example, there are many things I need to do to actually vote: I need to see to it that I get registered to vote, I need to find time during the day to get free from competing obligations, I have to find out where the polling place is and get there. Michael Bratman (1999) argues that intentions are best seen as something like partially specified plans that serve to regulate our further deliberations and choices as to what actions to take. If this is right, there is often a big gap between intentions and actions and a causal story seems implausible.

Second, we need to ask how intentions are formed. Some kinds of intentions might arise causally: if you haven't had any liquid in couple of days and a glass of water appears in front of you, you might "automatically" form an intention to drink it. But most intentions do not arise automatically in that way. On the deliberative account, the intention to do X arises from X's being the best thing for you to do in circumstance C. But being the best thing to do is a normative notion and so seeing X (including its intentional aspect) as caused seems to involve saying something like that norms have causal force. I agree that they have force, but of a different kind.

I think it is more natural to work the other way around and to try to assimilate external to internal explanation. That is, perhaps we should bite the bullet and admit that human action is deliberative and that we cannot escape the necessity of recognizing the priority of an internal perspective if we want to understand action. In particular, I want to argue that the notion of causation in human affairs needs to be removed from the physicalist model and given an understanding that fits with internalism. This will involve restricting the role that causal explanation can play in explaining human action. This may seem like a strange thing to do but I think there is a familiar model for making exactly this move: rational choice theory, or at least what I propose as the best way to understand the relation of that theory to actual human behavior.[7]

[7] For purposes of this chapter I am not going to keep reminding the reader that there are many variants of rational choice theory, indexed largely by the preferences that agents are assumed to have. Some variants of the theory posit that people are motivated to maximize their wealth, or perhaps their social status, or in some other more or less private-regarding way. Other versions, more common in political science, posit ideological motivation. And still others make no substantive assumption as to what ends agents desire – only that the desires, beliefs, and actions are connected in the way the rationality hypothesis requires. Sometimes this distinction is summarized by saying that the latter agents are "thin rational" and the others "thickly rational." This is an unfortunate usage in that it suggests a continuum or dimension of thickness, whereas like Tolstoy's unhappy families there are many different kinds of thickly rational agents, and therefore many different thick rational choice theories.

V. A social ontology of rational choice theory

Rational choice theory can be understood to offer both internal and external accounts of action. Internally, rationality explanations work through a mental model in which beliefs, desires, and actions are supposed normally to be related to each other. This is an internal idea if we think of this relation being established deliberatively, in reasoning about what to do, or think, or believe. But externally, rational choice accounts seem to rest on some presumed causal regularities as well. Having beliefs B and preferences P can be understood to cause act A where A is a best action given B and P. There are some problems for the external interpretation of course: what if A is not the unique best act? How do we understand the failure of actors to take best actions, or to exhibit variability in their choices? From an externalist viewpoint we can solve these problems pragmatically (by representing decision problems in such a way that there are always unique best responses, or by taking beliefs, preferences, and actions to be random variables whose values are determined according to stochastic assumptions, or by recognizing measurement error, etc.).

Even with these assumptions in place, however, I don't think that the external view of rational choice theory is compelling. It makes causal claims that simply seem to deny that actual agents make choices and could refuse to make them as well. At best, the external account offers a causal story about perfectly rational machines that are in fact not deliberating at all but are infallibly picking best actions given their "preferences" and "beliefs," together with a claim that real human actors will behave in approximately the same way. I shall argue that there is a plausible way to put this story. It entails seeing the rational machines of the previous sentence as machines that are designed to implement a norm of rationality. The machines will behave rationally according to some causal processes – which depend on how they are built internally, whether out of gears and levers, digital computing devices, or whatever – because they are designed to implement a rationality norm. From the standpoint of the norm, the connection among B, P, and A is of course not causal

but is conceptual or, if you prefer, constitutive. Choosing best acts given beliefs and preferences is what it means to be rational; we would not call our machine rational unless it actually did implement the rationality norm.

Putting the matter this way also illustrates the connection between internal and external explanations: real human agents regard rationality as normative – they take it as a defect of their action choices that they fail to fit with their beliefs and desires – and that is why internal accounts are illuminating.[8] And, the fact that real human beings are effective enough in actually behaving as rationality requires gives us reason to think that we can explain some of their behavior in the external way. This metaphorical talk of machines needs now to be grounded in human behavior. Indeed, if humans were not successful in this way, it is hard to see how social life would even be possible. So at least as a rough statistical matter, we can understand much of what others do, by taking an external explanatory "stance" with respect to their behavior. But, it must be remembered that a stance or perspective – even a highly successful one – is not the same thing as an ontological claim about how humans act.

Pettit (1993) has argued that human beings are not merely intentional systems (entities that seek ends), though they are that, but are also thinking beings.[9] So, satisfactory social explanations – and here he means internal explanations – need to take account of both aspects of humanity. The notion of an intentional subject is familiar enough not to need discussion; what is new in Pettit's idea is that humans have a distinctive capacity (thinking) that distinguishes them from merely intentional systems (animals, thermostats, etc.). He argues that this capacity – perhaps the most distinctive such capacity – is the capacity to identify and follow rules. By this he means the capacity to identify and follow rules – of the kind Wittgenstein, Kripke, and others have discussed – that are potentially applicable in indefinitely many circumstances. Philosophers writing on this topic typically choose their examples from language or mathematics, and are usually concerned to show that rule identification is impossible. But as Pettit argues, these examples support the idea that

[8] This statement needs more defense and qualification than I can offer here. I suppose actually that agents do not really take formal rationality as a norm directly, but are instead disposed to act in ways that further their aims. Effective pursuit of aims, however, entails that the agent is generally acting rationally in a formal sense. So, I would argue that the norms that agents accept are concretely related to the aims they actually have, and the alternative actions they may actually take.

[9] My account differs somewhat from Pettit's in that I argue that rationality is a norm and not a property of human beings as intentional systems. Moreover, I think that the rationality norm has the characteristics of a rule that actually requires thinking to follow.

the capacity to follow rules, if there is one, must rely in some way on background assumptions and cannot be infallible. Norms may be taken to be special examples of rules in that they are supposed to apply in indefinitely many situations and to direct action in those cases. Therefore, on this account, the capacity to be guided by norms is an instance of the capacity to identify and follow rules more generally. Moreover, the rationality hypothesis, or any variant of it, may also be seen as a rule in Pettit's sense, in that it directs the choice of action in indefinitely many circumstances and conforming to rationality hypothesis is an example of rule following. Thus, if Pettit's conception of rule following is right, we might expect human beings to exhibit (at most) imperfect rationality, in the same way that they would be capable of complying (imperfectly) with other norms and rules.

On this account we can understand that human beings are physical entities, subject to ordinary causal laws, and that these causal regularities produce in them certain (more or less hard-wired) capacities to act as particular norms direct (however imperfectly). One could well ask why physical human beings would be impelled to conform to a norm even if they could do so. This is a question of motivation and is a familiar problem for moral theories. Philosophers since Aristotle have insisted that seeing what one ought to do is one thing, and being motivated to do it is something else altogether. Indeed, a common criticism of a moral theory is that it fails a test of motivation: that it makes demands on people that they cannot ordinarily be motivated to respond to. Various versions of utilitarianism are often criticized in this way.

Conversely, some versions of rational choice theory are supposed to offer particularly compelling reasons to people. Material self-interest, for example, is thought to be a strong motivation in many circumstances (one can imagine evolutionary causal mechanisms that might explain why this would be so). If this is right, one would expect the behavior of physical humans to conform roughly to the operation of that version of the rationality hypothesis, at least where there are not other competing and compelling normative reasons to refrain from such action. One can also imagine causal mechanisms that would support status motivation and many biologists think that one could similarly explain the attraction of certain forms of altruism as well.

So, there is a role for ordinary causal processes in this story; ordinary causal mechanisms will no doubt play a part in shaping human capacities. And, the attraction of a norm to an agent may also be explained causally. But there is no asserted causal relation among the mental attitudes of the human beings and their actions. The relation among these entities

is tautological (with beliefs B and preferences P, a rational entity would, definitionally do A),[10] and normative (doing A is attractive to the agent on the grounds that this is what rationality recommends). In this respect, the rationality hypothesis offers an internal account of action: rational agents choose best actions, based on the reasons they have, and so their choices are justifiable. But it offers an external account of action insofar as capacities and motivations play a causal role in setting up the parameters for rational deliberation.

This view is, it seems to me, quite different from the standard interpretation of rational choice theory, one that seems accepted by both proponents and critics. On that account, rationality can be understood as a property of the human mental or neural apparatus that generates behavior causally. Put a (fully described) rational creature in a certain circumstance and it automatically does a specific thing. As I suggested before, this view seems (needlessly) to deny that people make choices. The construction I offered above leaves room for deliberative rationality and choice while, at the same time, recognizing some room for causation (in forming capacities and motivations). But the causal aspects of an explanation have little to do with rationality itself. In that sense the main work of rational choice *theory* is internal. Whether there are, additionally, external/causal implications turns on other facts about humans (capacities and motivations).

One implication is that the propositions from rational choice theory cannot be expected to represent or approximate causal regularities exhibited in human action. They are statements of how rational creatures would act – they are statements about how normatively directed creatures would act and statements about how people trying to follow that norm should act – and whether actual people come near to attaining that standard is going to depend on how closely they either can or want to comply with the norm itself. I have little to say about the capacity to follow rules (including rationality) in general or how that capacity might vary over contexts. But the attractiveness of various rationality norms seems likely to vary significantly. Consider, for example, the wealth-maximizing

[10] I won't go into the matter here but there is a deceptive simplification in this formulaic expression of the rationality hypothesis. Beliefs and preferences are both normative objects too. There are things, based on what I have experienced, that I ought to believe. No doubt my actual beliefs only approximate to the beliefs I should have. Perhaps more controversially, I think the same is true of preferences. There are things I should want and perhaps I do not want them at present. There are various ways to make sense of this; perhaps I don't fully appreciate the need for a gas-driven generator because I don't assign a high enough probability to a power failure, or because I don't fully appreciate what it would be like not be able to run my expresso machine in the event of a power outage. For now, these issues are off the table.

variant of rationality. Such a norm is likely to conflict with other attractive norms in various contexts. An agent might, for example, find wealth maximization an attractive guide where she is deciding on a retirement portfolio. But, she may resist following that norm if it recommends purchasing shares of tobacco firms. So we might think that any particular version of the rationality hypothesis will work better in domains without attractive competing norms.

VI. Limitations of rationality: a digression

I should add here that I am taking rationality explanations as examples of social scientific explanations in that they purport to offer causal or positivistic explanations. I do not deny that other "paradigms" may offer other positivist accounts of behavior. But, a purely causal explanation of human action seems defective on grounds I argued above; a theory that fails to recognize and represent how things seem from the agent's viewpoint, deliberatively, is missing something important about human action.

As an example, bounded rationality theories purport to offer alternative more or less causal or external accounts of human behavior, perhaps sometimes better accounts in some contexts than explanations offered from a rational choice perspective. Indeed, the rule following account offered here might seem, on the surface, to offer a reason to believe that as causal explanation, bounded rationality would likely be superior to rational choice theory. After all, bounded rationality stories seem to rely on what humans, as they are hard wired, are capable of doing. Such stories emphasize the incapacities of people to be rational even when they want to. And, on my account, these incapacities are likely to be the kind of thing that can be explained causally. So we might very well expect such theories to lead eventually to superior causal explanations of human behavior. But will such theories offer better accounts of human action – of what people do intentionally or deliberatively? Here I have doubts.

Can there be norms of bounded rationality? On the surface there seems to be no problem. Bounded rationality has the outward appearance of a norm: it directs an agent to do or refrain from doing things in indefinitely many circumstances. But, there are some internal problems too. Assuming, as I do, that we are cognitively constrained, we can nevertheless imagine, or build, systems capable of more rationality than we have. That is, at least in specific domains, we can build systems that transcend our limitations or capacities to behave rationally. We buy spreadsheet programs and programs to compute our taxes, presumably because, within

their task domains, they can be counted on to do a better job at complying with rationality requirements than we would do ourselves. Expert chess computers can by now beat all human players, and are presumably less cognitively constrained on chess problems than we are. Indeed, long ago it was proved that fully rational strategies exist in chess (or at least in the near relatives of chess that have finite game trees; i.e., the versions of chess that are played in professional chess tournaments).

So, at least in these cases, we could imagine a norm of bounded rationality but probably few of us would find such a norm compelling. We want to find out the best or most powerful chess program, not one that best tracks our limited capacities. We remain dissatisfied if we are told that there are better answers available even if we ourselves cannot reliably find them. So, in this sense, a boundedly rational "norm" has some real defects from an internal perspective; we know it would be giving us wrong or inadequate advice. It may represent the best that we can do on our own behalf, being wired as we are. And, indeed insofar as bounded rationality theories are inductively derived from empirical experience, their appeal is pretty much purely external.

From an explanatory standpoint, the attraction of bounded rational theories is that they provide a wedge by which recognizably causal factors can come into play directly in explaining human behavior. We can say that an actor took the action she did in part because she was cognitively unable to recognize the existence of better alternatives, or even to see what she did as an occasion to "choose" at all. Perhaps those statements are parts of good explanations. But then, have we really made her action intelligible? We may have explained why she failed to see a choice that was really there but, at the same time, we have made the action she took less than a full-blooded deliberatively chosen action. That may be the best route to a more or less causal account – perhaps what we initially thought was a deliberate action was actually something less than that. But, in some respects, adopting this explanatory strategy involves abandoning the idea of explaining human action rather than mere behavior. Explaining action – deliberatively chosen behaviors – requires something more than this. It requires showing that what the agent did was a best or most effective way of pursuing her purposes. And this entails establishing the embedded normative assertion.

VII. Conclusions

Rational choice theory provides one way in which the internal and external perspectives can be bridged. And the hermeneutical perspectives

taken by ethnographers exemplify another way of understanding these perspectives. Obviously, there are very important ways that practices of descriptive ethnography and theoretical economics differ, but they share one important feature. Both are, at bottom, normative enterprises aimed at showing how the agents – real or imagined – can act and have reason to act in ways that comply with norms they accept. Both also make more or less implicit causal claims that agents will generally tend to act as norms they accept direct them to. But this causal claim is not really a part of either theoretical apparatus. That the causal assertions tend, in many circumstances, to be empirically true makes these approaches useful to understanding social action.[11]

Naturally enough, these two bridging paradigms have generated distinct methodological programs and are fitted for addressing distinct explanatory problems. It is too easy, however, to overstate these differences and to lose the sense in which the common subject matter – understanding human activity – yokes these perspectives. Most often, the real differences between these approaches is driven by practical considerations that arise from the explanatory questions that are asked rather than any deep division between them.

Methodology, on this account, is fundamental to the social sciences. Methodological awareness forces us to recognize the pluralism of social reality: the fact that description and explanation are relative to perspective; that human beings are embedded in causal processes but are also responsive to and guided by norms. So a self-conscious focus on methods cannot be abandoned without giving up on social science, and indeed history, in favor of mere uncritical chronicle.

REFERENCES

Allison, Graham. 1971. *The Essence of Decision*. Boston: Little, Brown, and Company.

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

Bratman, Michael. 1999. *Faces of Intention: Selected Essays on Intention and Agency*. Cambridge University Press.

Durkheim, Emile. 1966. *Suicide: A Study in Sociology*. John A. Spaulding and George Simpson (trans.). New York: Free Press.

[11] Debra Satz and I have argued elsewhere that different causal processes might explain why and when agents respond to normative recommendations that arise from rationality. Some of these might be internal to the agent, some might be learned, and some might arise from the environment of interaction. But whatever it is that explains when rational choice theory provides a successful account of human action is itself part of rational choice theory. See Satz and Ferejohn (1994).

Fenno, Richard. 1978. *Homestyle: House Members in their Districts.* Boston: Little, Brown and Company.
Krasner, Stephen D. 1972. "Are Bureaucracies Important? (Or Allison Wonderland)." *Foreign Policy* 7: 159–79.
Pettit, X. 1993. *Common Mind.* Oxford University Press.
Satz, Debra and John Ferejohn. 1994. "Rational Choice and Social Theory." *Journal of Philosophy* 91: 71–87.

Part II

Redeeming rational choice theory?