RATIONAL CHOICE THEORY AND EXPLANATION

Frank Lovett

ABSTRACT

Much of the debate concerning rational choice theory (RCT) is fruitless because many people (both critics and defenders) fail to correctly understand the role it plays in developing explanations of social phenomena. For the most part, people view rational choice theory as a species of intentional explanation; on the best available understanding, however, it should be viewed as contributing to the construction of straightforward causal explanations. Debate concerning RCT can progress in a worthwhile manner only once this point is correctly understood. Once it is, many common critiques are easily answered, but at the same time, the ambitions of some rational choice theorists are deflated.

KEY WORDS • game theory • methodological individualism • social science epistemology • utility theory

Introduction

The literature attacking and defending rational choice theory (RCT) is enormous, and still continues to grow; debating its merits, or lack thereof, has become something of a *cause célèbre* among social and political theorists. Much of this debate proves fruitless, however, because many people (both critics and defenders) fail to correctly understand, or at least fail to sufficiently appreciate, the role rational choice theory plays in developing explanations of social phenomena.

Empirical explanations can usefully be divided, following Jon Elster (1979: viii, 1983a: chs 1–3), into three basic modes: causal,

Rationality and Society Copyright © 2006 Sage Publications. Vol. 18(2): 237–272.
www.sagepublications.com DOI: 10.1177/1043463106060155

functional, and intentional.[1] In its standard form, the first explains a social phenomenon by referring to an antecedent event or state of affairs that, together with some sort of causal mechanism, is sufficient to bring it about deterministically.[2] The second explains a social phenomenon teleologically – that is, by referring to the purpose or function it serves. When successful, functional explanations may arguably be a special case of causal explanations, though their distinctiveness warrants viewing them as separate.[3] The third explains a social phenomenon as arising ultimately from the intentional states – desires and beliefs – of social actors (together, perhaps, with some intervening chain of causal or functional explanations carrying us from these decisions to the outcome in which we are interested). If it could be shown that intentional states were causally sufficient antecedent conditions for human action, then the intentional mode of explanation might also collapse into the causal mode.[4] There are important philosophical reasons for doubting whether this could be shown, however, and in any case my article will remain agnostic on such questions.[5]

For the most part, people view RCT (which will be defined more precisely below) as a species of intentional explanation. Roughly, according to this view, the point of RCT is to explain social phenomena by showing how they arise from the deliberate or intentional pursuit of self-interest by social actors (and especially, individual persons).[6] These explanations can be more or less direct. They are more direct when social actors simply choose those outcomes that are in fact optimal for them, given the constraints they happen to face; they are less direct when, for example, outcomes are the unintended and perhaps undesired by-products of attempts to optimize in this manner.

Less often, people view RCT (or at least certain applications of it) as a species of functional explanation. In this view, the point of RCT is to show that certain social phenomena can be explained with reference to their usefulness in solving problems arising from the general pursuit of self-interest. For example, promise-keeping has the structure of a prisoners' dilemma, and so it might be difficult for a community to maintain promise-keeping as a social convention; the fact that state-enforced contract law can to some extent mitigate these difficulties might then be regarded as explaining (functionally) those legal institutions. This application of rational choice theory is common in the so-called law and economics literature.

Although nearly everyone views RCT as a species of either intentional explanation (more often), or else functional explanation (less often), it ought to be viewed as neither. Rather, on the best available understanding, RCT should be viewed as contributing to the construction of straightforward causal explanations. The debate over RCT has been fruitless because those contributing to it generally fail to understand RCT in this way.[7] Because the intentional explanation view is the dominant one, my article will concentrate on showing that it is inferior to the causal explanation view, and more or less ignore the functional explanation view.[8] I might also note that, since David Hume at least, the viability of causal explanations themselves has been the subject of extensive philosophical controversy, and no attempt will be made here to address this debate. Though fascinating, such questions are beyond the scope of this article.

Whether one regards this article as defending or attacking RCT depends on what one wants or expects RCT to do. Insofar as many common critiques are easily deflected once RCT is correctly understood, my article may be seen as defending the rational choice approach. But some want RCT to be something more than (as described here) a mere contribution to the development of straightforward causal explanations, and thus would regard the rational choice program as a failure if it did not succeed in doing more. From this point of view, my article may be regarded as an attack on RCT. It is not important which point of view the reader adopts in this respect; what is important is that people better understand the role of RCT in constructing social-science explanations. This is the only way, it seems to me, for the rational choice debate to progress in a worthwhile manner. In other words, this article primarily aims to clarify, rather than resolve, the rational choice debate.

The discussion is divided into two main parts. The first is concerned with laying out an account of RCT; its main goal is to convince the reader that rational choice models are best understood as contributing to the construction of ordinary causal explanations, rather than as providing intentional explanations. The second part is concerned with showing how this better understanding of rational choice theory effects our view of some standard disputes in the rational choice debate.

Rational Choice Theory as Causal Modelling

I

What do we mean by ‘rational choice theory’? Naturally, people disagree. The simplest and best answer defines RCT as an approach to the study of social phenomena characterized by a small bundle of core methodological assumptions. The core assumptions in question are three.[9]

First, there is what might be called a *discrete purposeful actor* assumption. This assumption maintains that there exist in the world of social phenomena discrete entities capable of acting purposefully. Human beings are, at least on the common-sense view of things, obvious examples of discrete purposeful actors. That is, human beings are discrete entities capable of considering several different possible courses of action, and deliberately selecting and carrying out (or attempting to carry out) one or more of them. Note, however, what the discrete purposeful actor assumption does *not* require: First, it does not require believing that individual human beings are the *only* possible discrete purposeful actors. Collective agents might qualify under certain conditions, for example. Second, it does not require believing that human beings *always* act in a purposeful way, only that they are capable of doing so at least some of the time. And third, it does not require believing that purposeful action is uninfluenced or unconstrained by factors external to the actor in question, so long as such influences and constraints leave the actor some choices, at least some of the time. Thus stated, the discrete purposeful actor assumption is so weak it is hard to imagine grounds for disagreeing with it. Often, the discrete purposeful actor assumption is confused with a different idea, commonly referred to as ‘methodological individualism’, which is more vulnerable to criticism. In part 2, I will argue that RCT does not depend on methodological individualism.

The second and third assumptions of rational choice theory are perhaps more familiar: these are the *utility theory* assumption and the *rationality* assumption. Under certain very general conditions (adumbrated below) it has been shown that an agent’s choices or decisions can be described as if they were an attempt to optimize a mathematical function, regardless of what they are actually an attempt to do.[10] The utility theory assumption holds that we can often expect the choices or decisions of discrete purposeful actors to conform with these conditions. When this is the case, it follows that we can assign each discrete purposeful actor a 'utility function', which provides, as it were, a concise mathematical summary of whatever choices or decisions we expect them to make. The rationality assumption is tightly related to utility theory. Roughly speaking, it states that we can expect discrete purposeful actors to optimize their utility functions, given whatever constraints they happen to face.

Since utility functions are merely a mathematical representation of what we expect discrete purposeful actors to do in the first place, these two assumptions are, in a sense, merely two sides of the same coin. Indeed, in strictly parametric situations the latter collapses into the former. However, when what one actor wants to do depends in part on what others might do, things get more complicated, necessitating a distinction between the two assumptions. For example, in ordinary market competition, each firm seeks to maximize its profits, and thus the specification of its utility function is elementary. What pricing strategy will actually maximize a firm's profits, however, depends in part on the pricing strategy adopted by competing firms. Thus it is not obvious, on the basis of their utility functions alone, what the rational pricing strategy for each firm would be – or at least not without further elaboration of what it means to behave 'rationally'. The standard elaboration employed in such strategic situations is provided by what is called the Nash equilibrium solution concept.[11] Roughly speaking (the details are not important for this article), the idea is to find a profile of strategies such that no one actor could do better by unilaterally changing her strategy, supposing the other actors do not change theirs. In the market competition example, this might suggest that each firm, behaving rationally, will continually attempt to undercut the prices of the others until they reach the point where prices equal production costs, and thus no individual firm can do better by either raising or lowering its prices unilaterally. In settling on this equilibrium, firms behave according to the rules of 'strategic rationality'. Because the rationality assumption embraces not only simple utility maximization, but also auxiliary principles like the Nash equilibrium solution concept for more complex situations, it is not, strictly speaking, reducible to the utility theory assumption.[12]

The utility theory and rationality assumptions are commonly confused with something quite different – what philosophers call the doctrine of ‘psychological egoism’. This confusion will be explained and discussed in the second part of the article.

These three methodological assumptions tie together the disparate research projects loosely grouped under the heading of RCT. For the sake of completeness, I should define two other terms often associated and sometimes thought interchangeable with rational choice theory: namely, *game theory* and *social choice theory*. Game theory is a loose bundle of modeling tools, at present widely regarded as the best tools for implementing rational choice research projects. It has been discovered, however, that game-theoretic modeling tools can be usefully employed in other areas of research as well (in evolutionary biology, for example), and so we must distinguish these from RCT as such. Social choice theory offers a somewhat different loose bundle of modeling tools; like the tools of game theory, these have some application outside the domain of rational choice theory, and so they too must be kept distinct. In what follows, the discussion will focus on game-theoretic rational choice models. Of course, game theory is not needed to model some problems, but these can always be interpreted as degenerate game-theoretic situations. Thus, while not every interesting rational choice problem *requires* the tools of game theory, game theory can model every interesting rational choice problem at least as well as any alternative technique.

Guided by the three methodological assumptions mentioned, rational choice theorists study social processes involving *constrained purposeful action*. There is no reason a priori to believe that all social phenomena can ultimately be explained with reference to processes of this sort. In some situations a person might face constraints so restrictive that she has no real choices to make. Supposing cases of this sort exist, RCT would contribute little or nothing to our understanding of them (unless, of course, those very constraints were themselves brought about by constrained purposeful action). In some situations, people do not make decisions purposefully, as for example when they decide what to do randomly; again, RCT would be of little help. Sometimes what is interesting about a social phenomenon is something other than the processes of constrained purposeful action that bring it about. Consider education policy, for instance. On the one hand, the design and implementation of education policy will involve a great deal of constrained purposeful action on the part of many social actors (school boards, legislatures, teachers, administrators, voters, etc.); these processes are in principle amenable to rational choice analysis. But it may happen, on the other hand, that what really interests us is rather the socio-psychological effects such policies have in shaping the beliefs, attitudes, dispositions, and so on, of children. These latter processes do not primarily involve constrained purposeful action, and so remain beyond the scope of RCT. Obviously, a modeling technique will not be good at modeling situations it is not designed to model, and there is no reason to believe rational choice theory is an exception: it is not, nor does it have to be, 'a theory of everything'.[13] I will return to this point in my conclusion.

II

Having defined RCT with sufficient clarity, I hope, we may now turn to the main topic of the article – namely, how we are to understand the role of RCT in social science. To gain purchase on this question, it might be useful to look at an exemplary example of rational choice research, and for this purpose I will discuss a model loosely based on Gary Cox's (1997) *Making Votes Count*.

At this juncture, however, some readers may wonder what sort of account of RCT I claim to provide: Do I only mean to describe what rational choice theorists are actually doing? Or do I mean to tell rational choice theorists what they should be doing? Or something else? This essay is clearly meant to be a contribution to the philosophy of science, but the philosophy of science is a peculiar sort of enterprise. It is not exactly descriptive, on the one hand, in the sense that it does not merely aim to provide a sociologically or anthropologically correct description of what scientists in any particular area of research are actually doing (or even a description of what they believe they are doing). Rather, it offers a philosophical reconstruction of some area of scientific research. In other words, it is primarily concerned with answering the question: What is the most philosophically sound account of what scientists (in a particular area of research) are doing? Actual scientists need not agree with the answer given, nor even be particularly interested in answering the question. In fact, they will sometimes believe they are doing something quite different, and so would not recognize themselves – at least not at first – in the portrait constructed of their enterprise by philosophers of science.

On the other hand, the philosophy of science is not exactly proscriptive either. Although the most philosophically sound account of a particular area of scientific research may not correspond perfectly with what practicing scientists actually believe they are doing, the point is not, in the main, to get scientists to start doing something else. The philosophical reconstruction remains a reconstruction of the practice scientists are actually engaged in, and not some hypothetical outline for some other practice they might turn to instead. Given a particular philosophical account of an area of scientific research, different examples of research in that area will seem more or less exemplary; at the same time, one's view of the best philosophical reconstruction will depend to some extent on which examples of actual research one takes to be more or less exemplary. There is, in other words, a curious circularity to the philosophy of science – though not, fortunately, a fatal one. On the contrary, the philosophy of science stands in a position vis-à-vis the actual practice of science much as Ronald Dworkin (1986: 90) argues the philosophy of law stands vis-à-vis the actual practice of law: legal philosophers, in his view, 'try to show legal practice as a whole in its best light, to achieve equilibrium between legal practice as they find it and the best justification for that practice'.

With these caveats in mind, let us turn to our exemplary example of rational choice research. Our starting point is an empirical regularity long known to political scientists, and referred to as Duverger's Law. Roughly, this holds that electoral systems relying on a simple-majority single-ballot system tend to support only two effective political parties in the long run, whereas those systems relying on proportional representation tend to support multiple parties. Although generally confirmed as an empirical regularity, the actual mechanisms underlying are not well understood. In order to better understand those mechanisms, Cox develops a simple game-theoretic rational choice model, which I loosely describe in what follows.

Game theory models include three main parts: first, a list of players; second, a specification of the rules of the game; and third, a schedule of utility functions. Beginning with the first of these, a list of players is merely the list of the discrete purposeful actors – be they individuals, firms, political parties, states, or something else – relevant from the point of view of the model. Roughly speaking, the only relevant actors are those facing choices or decisions within the scope of the social phenomena one is interested in better understanding. Cox is not, for example, interested in electoral corruption. From the point of view of his model, therefore, we can treat the behavior of ballot-counters as automatic, and so we need not include them in our list of players. The most obvious candidates for inclusion are the voters themselves, insofar as they *do* face relevant choices – namely, how to cast their ballots. Other actors (political parties, interest groups, etc.) might also be relevant, and so should not be excluded a priori; but, as we shall see, a simple model with voters alone turns out to be sufficient.

Turning next to the game's rules, we must consider the different possible paths of action open to our players. A path of action is by convention called a 'strategy'. Complete strategies might involve a series of actions, some of which are contingent on future events. For example, a congressman's strategy in some game might be 'vote in favor of the amendment; if the amendment passes, vote in favor of the bill, and if the amendment fails, vote against the bill'. Another might be 'cooperate with congresswoman Jane today; the next day, cooperate again if she cooperated the day before, otherwise not; the day after that cooperate again if she cooperated on the second day, otherwise not; and so on'. The different strategies available to a given player depend on constraints such as the laws of nature, social or cultural conventions, economic forces, personal endowments, and so on. Some would call these constraints 'structures', but game theorists call them the 'rules of the game'. The rules of the game offer each player her own set of possible strategies, called a 'strategy space'. Strategy spaces and rules of the game are exactly the same thing, described from two different points of view, the former emphasizing what a player can do, the latter what she cannot do. In the Cox model, the strategy space for each player (voter) is the same: it is simply the set of possible ballots each can cast in an election. It might include, for example, 'vote for candidate or party list *A*', 'vote for candidate or party list *B*', and so on.

Finally, we come to the schedule of utility functions. Because the assumptions of utility theory are the source of such controversy, it might be useful to go into somewhat greater detail here. In particular, it is important to appreciate what utility functions are defined over in a game-theoretic rational choice model. Let me explain.

The strategy space for each player, as we have said, represents the various paths of action open to each. Different strategy combinations lead to different outcomes. Imagine a game with two players, and suppose that player 1 must decide between three paths of action, while player 2 must decide between two (see Figure 1). The strategies here have been given dummy names. Here we see that each combination of strategies – what is called a 'strategy profile' –

		Player 2		
		Left	Center	Right
Player 1	Top	x_1	x_2	x_3
	Bottom	x_4	x_5	x_6

Figure 1.

leads to a different outcome. For example, the strategy profile $\mathbf{s} = (\text{left, top})$ leads to the outcome x_1.[14] The set of possible outcomes, in this case $X = \{x_1, x_2, x_3, x_4, x_5, x_6\}$, is called an 'outcome space'. When there are more than two players, it becomes difficult to express strategy profiles in two-dimensional figures like this one. In Cox's model, for example, if there are three voters and each can vote for one of three candidates, there would be $3 \times 3 \times 3 = 27$ possible outcomes.

A utility function is defined over the outcome space of a game. Suppose we ask player 1 a series of questions of the sort, 'If you could choose, would you prefer outcome x_1, or x_2, or are you indifferent between them?' It does not matter *why* she prefers one to the other; her reason might be perfectly altruistic, for example. Nor does it matter whether the outcomes are different in terms of the player's material well-being (unless, of course, this matters to the player): even if player 1 is equally well off with respect to outcomes x_1 and x_2, she might nevertheless care how that level of well-being came about. For example, she might dislike the thought of player 2's having played 'center', and so prefer x_1 to x_2 for *that* reason. Since a fully-specified outcome space includes all possible histories of the game, nearly anything at all relevant from a player's point of view can be captured in a utility function defined over that space. And provided the players' expected choices conform to the rules of utility theory – for example, that their ranking of outcomes be internally consistent – it will always be possible to construct one.[15] The utility functions we assign them, then, are just those functions we expect them to act *as if* they were trying to maximize.

To be perfectly clear, no one should suppose there really exists some such thing as 'utility' which people actually go around trying to maximize. The utility function is merely an artifact of the model – a compact mathematical description of the outcomes a player would choose to bring about, given the opportunity to do so. Real social agents ‘have’ utility functions only in the sense that there (usually) exists at least one mathematical function correctly representing whatever it is we expect them to do. Indeed, there are typically many more than one. Putting aside uncertainty for the moment, to say that $u_1(x_1) = 9$ and $u_1(x_2) = 5$ means only that we expect player 1 to bring about outcome x_1 rather than outcome x_2 given the opportunity to do so, and saying $u_1(x_1) = 1003$ and $u_1(x_2) = 0.074$ conveys exactly the same information – i.e., it is equivalent in terms of describing what it is we expect her to do. To ask which of these is the ‘correct’ function is to commit the fallacy of misplaced concreteness: when several utility functions capture our behavioral expectations equally well, we are free to choose the one most convenient for our model.

Now the outcome space in a model like Cox’s is quite large indeed: specifically, it has v^n possible outcomes, where n is the number of voters, and v is the number of voting options available to each voter. For example, if nine voters must choose one of two candidates, there are, strictly speaking, $2^9 = 512$ possible outcomes. Each player’s utility function must be defined over this space. A player might care only which candidate wins, in which case we could assign half those possible outcomes a utility of 0 and the other half a utility of 1. Alternatively, she might care how many votes a candidate receives, in which case the nine possible outcomes in which her preferred candidate receives one vote would all have the same utility. Or again, she might care *who* votes for which candidate: perhaps she might be happier, even when the vote totals are the same, that her candidate did not receive a vote from player 4. For the purposes of the Cox model, we can make the simplifying assumption that voters care only which candidate or party wins, and so the player’s utility functions need only rank-order the candidates or parties contending in a given election.

Nor is this the only simplifying assumption we make. It is assumed that no other information need be included in the players’ utility functions. It is assumed that each player can rank-order *all* the contending candidates or parties *consistently*. It is assumed that every player will vote. More importantly, it is assumed that each player goes to the polls with well-defined beliefs regarding the likely electoral outcome – i.e. that most people prefer candidate *A*, that *B* and C will probably tie for second, *D* will be a distant third, and so on. And of course it is assumed that each voter will behave rationally, meaning in this case that each will cast her ballot so as best to achieve her desired electoral outcome, given what she expects other voters to do. Obviously, each of these assumptions is open to challenge. Indeed, some of them are clearly false. But does this matter? Would a model with more accurate assumptions tell us more than this (admittedly) crude and simplistic model? To answer this question, we must consider what it is we expect the model to do.

III

What do rational choice models tell us about social phenomena? Suppose we observe a relatively stable pattern of social behavior that is somewhat puzzling. Game theory offers us a set of tools that can help us understand how this stable pattern of social behavior might arise, given some prior configuration of structural constraints we already know something about. Thus, like any other sort of explanation, a game-theoretic rational choice model explains the less familiar with reference to the more familiar. But game theory has some especially desirable properties that other sorts of explanation do not have: first, it does not black-box the causal process leading from the less familiar to more familiar, but on the contrary draws it out as explicitly as one might reasonably hope; second, it does this using a conceptually simple, yet also extraordinarily flexible set of tools shared by nearly all rational choice theorists.

However, this very flexibility itself can be puzzling, for it is, in a sense, *too easy* to explain a stable pattern of social behavior using the tools of game theory. Indeed, under a fairly wide range of conditions, it is virtually guaranteed that some rational choice model (and usually more than one) can be constructed for the situation one happens to be interested in. Often, it is simply a matter of tweaking the rules of the game and the players' utility schedules until the desired result is obtained. But if this is so, how can RCT ever be falsified? It cannot be. But it is not supposed to be the sort of thing that can: rather, RCT should be understood as an approach to developing models that can. In other words, the models are the things that can be falsified, not the approach itself. When a particular game-theoretic model fails to yield the correct result, one needs a better model, not necessarily a new approach.

So how are rational choice *models* falsified? This too can seem puzzling, insofar as it is usually possible to engineer a model so that it yields the result one wants. The key to dispelling this puzzle lies in distinguishing *point predictions* from *comparative statics*. Let me explain.

Having completely specified a model, like the one described in the previous section, we apply the principles of rationality to figure out what we expect to happen. If what each player wanted to do were perfectly independent of what the others did, it would be enough to say that each chooses the strategy leading to the outcome with the greatest utility for her. In our voting model, however, what each voter wants to do clearly depends on what she expects the others to do. Suppose for example that player 1 prefers candidate *A* to *B*, and candidate *B* to *C*. If she expects others to vote in such a way that candidate *A* has a shot at winning, she will naturally vote for *A* herself. But if she expects others to vote in such a way that *A* has little or no shot at winning, she might vote for *B* instead. Whether she will or not depends, of course, on the relevant odds and the strength of her preferences; the point here is only that the situation is a strategic one, in that what she wants to do depends in part on what she expects others to do. Thus we need the Nash equilibrium solution concept (or some substitute) in order to obtain a result.

Now suppose we have done this. What we have in hand – an equilibrium result in a game-theoretic model – is a point prediction. That is, we have a predicted specific outcome of the sort 'candidate *A* receives 52% of the votes, candidate *B* receives 40%, etc.' If all has gone well, this result in our model will match or correspond in some recognizable way to the stable pattern of social behavior we are actually interested in explaining. At the same time, it may be wildly inaccurate in many respects. The Cox model predicts that everyone votes, that no one votes expressively, that no one makes erroneous probability estimates, and so on – all of which is clearly not what we observe in the real world of social phenomena. By modifying the model's specifications to include further complexities we could, to be sure, engineer a result considerably closer to the one we actually observe, though we could never achieve an exact match. What is more, there might be several other game-theoretic models quite different from ours, all yielding equally good point predictions. Should it bother us that two or more models generate equally good predicted outcomes? Should we strive to make our model's predicted outcome as close as possible to what we actually observe, and should the impossibility of our achieving an exact fit bother us? If we were judging models on the basis of their point predictions, the answer would be *yes*. I submit, however, that it is not the point prediction of a model that is interesting, but rather its comparative statics. This can be seen with the help of a parallel example.

Models of voter turnout are widely assumed to represent a spectacular failure for rational choice theory. The difficulty is that they consistently predict levels of turnout far lower than what we generally observe.[16] A given level of turnout, however, is a point prediction. In itself, it is not very interesting, because with the addition of sufficient mathematical epicycles, a rational choice model yielding something like the 'correct' level of turnout could no doubt be constructed. What *is* interesting is the mechanism proposed: the models in question generally hypothesize that in deciding whether or not to vote, people will weigh the strength of their preference for a particular result and the probability of their vote making a difference on the one hand, against the voting costs imposed by the electoral system (registration hurdles, difficulty in getting to the polls, etc.) on the other. If this proposed mechanism is correct, then we would expect to see certain things. For example, we would expect that an increase in the cost of voting (say, adding a registration fee) will decrease turnout. This is an example of a comparative static. Comparative statics are determined by changing the parameters of the model and calculating how the predicted result changes. In this case we raise the cost parameter, and expected turnout in the model falls. This gives us a testable empirical prediction capable of falsification: if we find this relationship does not hold, then the model must be false.

It is irrelevant for the argument here what the data actually show. For the sake of argument, imagine someone examines the data and confirms the comparative static prediction. This lends credence to the model. What sort of explanation of voter turnout do we have then? In particular, do we have an *intentional* explanation? We do not. Intentionality entered the model only as a postulated mechanism connecting the independent variable (structurally imposed voting costs) with the dependent variable (turnout level). Since the intentional states of social actors were assumed and held constant, they were not subject to empirical test, and so whatever explanatory power the model has, it must lie elsewhere. What *was* tested, of course, was an ordinary causal explanation – namely, one that explains turnout with reference to the structurally imposed costs of voting, together with the intentional behavior of voters acting as a causal mechanism.[17]

Now let us return to the Cox model. It is not the model's point predictions – actual electoral results – that concern us. Thus, it does not matter that the assumptions are in many cases implausible, and that the equilibrium results are inaccurate. Rather, it is the comparative statics that are interesting and empirically testable. Two comparative statics in particular emerge from the model. The first is generated by altering the rules of the game – specifically, the rule by which raw ballots are converted into electoral results. One rule states that 'the candidate with the most votes wins'; another states that 'the two candidates with the most votes win'; still another states that 'seats are apportioned among lists of candidates according to the percentage of votes cast for each list'; and so on. The second comparative static is generated by shifting the players' beliefs and expectations. The effect of the latter has already been seen. Considering the voter described earlier who prefers candidate *A* to *B* and *B* to *C*, we saw that how she will cast her ballot will depend on her beliefs concerning the likely electoral result. If she does not expect *A* to have a shot at coming in first, she might vote for *B* instead. Shifting the players' expectations will thus alter the equilibrium result. But we have assumed that only the candidate with the most votes wins. If the electoral rule states instead that the two candidates with the most votes win, our voter might vote for *A* after all, provided that she believes he has at least some chance at coming in second. Thus we find that the Nash equilibrium result of our model is the joint product of the electoral rules on the one hand, and the voters' beliefs and expectations on the other. Varying either parameter varies the predicted result of the model. These two comparative statics are both interesting and relatively easy to test empirically.

What is more, our model – crude and simplistic though it may be – greatly improves our understanding of Duverger's Law. In particular, it transparently unpacks the causal mechanism leading from electoral rules and voter expectations on the one hand (the independent variables) to the effective number of political parties on the other (the dependent variable). When people vote strategically to maximize their desired electoral outcomes, simple-majority single-ballot systems inexorably compel them to vote for one of the two candidates most likely to succeed. In a single-ballot system where the top two candidates win, the exact same logic compels them to vote for their most preferred candidate among the three most likely to succeed. Indeed, we may now formulate Duverger's Law in a more precise and general manner. Suppose we define 'average district magnitude' (m) as the average number of candidates who can win in each electoral district. (Since the USA uses a simple-majority single-ballot system exclusively, $m = 1$. By contrast, in Israel there is only one electoral district – the whole country – for a legislature with 120 seats, so $m = 120$. Other systems fall on a range between these.) Stated precisely, then, Duverger's Law predicts the maximum number of effective political parties in any electoral system will tend towards $m + 1$ in the long run. And indeed, this is precisely what the data show (see Cox, 1997: esp. ch. 11).

As in the case of voter turnout models, let us consider what sort of explanation has been constructed here. On the standard view of RCT, the object must have been to construct an intentional explanation leading from the deliberate effort of individuals to maximize their expected utility to particular electoral outcomes. But if this is what the Cox model is supposed to do, surely it fails. The rational choice model of behavior was merely assumed, and never tested against a competing hypothesis. This is not to say that intentionality plays no role in the complete explanation of Duverger's Law, of course – only that it is not the part subject to empirical test. That the Nash equilibrium result of his model does indeed approximate the electoral outcomes we actually observe is neither here nor there, for different models might easily generate similar results. The substantive empirical content of the model lies in its comparative statics. In other words, by assuming X (the strategic behavior of voters) as a constant, the Cox model easily produces several empirically falsifiable tests of the claim that Y (electoral structure) causes Z (the number of effective parties). Moreover, it does this perfectly well *despite* being crude and simplistic, built on many assumptions known to be false.

Of course one might set out to do something quite different. One might, for example, start by trying to figure out what people care about and how much they care about it, and then examine the extent to which they actually behave intentionally so as to realize those goals or ends. If successful, this would indeed constitute an intentional explanation. Perhaps an example of such an attempt is the so-called 'attitudinal model' of judicial behavior, which argues that, contrary the standard view in legal circles, judges vote so as to best realize their ordinary political preferences.[18] But this counter-example only supports the general point, for confirming the attitudinal model is a statistical and not a game theoretic problem, nor would success or failure in doing so say anything about the rational choice approach generally. The dispute to which the attitudinal model is a contribution is basically a dispute concerning the shape of judges' utility functions, on which question RCT as such remains perfectly agnostic.[19] In actual practice, rational choice theorists for the most part use intentional mechanisms merely to construct and test causal explanations of the usual sort. Hence the main contention of this article – that on the best available understanding, RCT should not be seen as an attempt to explain social phenomena intentionally. Only having cleared up this point can we move on to a fair assessment of its value.

Critiques of Rational Choice Theory

In the second section of this article, I assess some common criticisms of RCT. For this purpose, it will be useful to provide a rough typology of the most common of these, comprising four main groups. First, there are critiques concerning the assumptions of utility theory. Second, there are related critiques directed not at utility theory itself, but rather at the rationality assumption. Third, there is a set of quite different critiques concerning the problem of what is called 'methodological individualism'. And finally, there are various critiques related to the practical execution of rational choice research projects. (Another cluster of issues, which I refer to as the 'endogeneity problem', will be discussed in the Appendix.) Considering each of these in turn, I will try to show how our revised understanding of rational choice theory affects our assessment of their seriousness.

I

Rational choice models rely on a number of utility theory assumptions. Some of these assumptions are mathematically arcane and uninteresting. Two important assumptions, however, are not: these are the consistency assumption, and the continuity assumption. Both will be explained shortly.

Before doing this, however, it is worth dismissing some common complaints that are the product of simple misunderstanding. These generally arise from the confusion of the utility theory assumption with something quite different – namely, what philosophers refer to as the doctrine of 'psychological egoism'. The doctrine of psychological egoism holds that human action is always and only motivated by the material self-interest of an actor herself.[20] Common complaints with psychological egoism are: that people sometimes act altruistically or on the basis of other-regarding motivations; that people sometimes act non-instrumentally because they value some activity for its own sake; that people sometimes act expressively or symbolically, without regard to the maximization of their interests; and so on.[21] These are all sound critiques of psychological egoism; none, however, contradict the requirements of utility theory. In other words, other-regarding, non-instrumental, expressive or symbolic behavior can perfectly well be represented *as if* it were an attempt to maximize some concise mathematical function.

At the risk of tedium, I will give an example in the case of other-regarding preferences. Suppose person i is motivated by the (other-regarding) desire to bring about social equality. Consider three possible distributions of goods between i and another person j, $x_1 = (9, 4)$, $x_2 = (5, 8)$ and $x_3 = (7, 6)$, where the first in each pair represents the distribution to i and the second the distribution to j. If, as supposed, i desires social equality, then she would rank these distributions $x_3 > x_2 > x_1$. In order to capture our expectations regarding her behavior, then, we should assign her a utility function that gives x_3 a greater real number value than x_2, and x_2 a greater real number value than x_1. A utility function that ranks x_1 first on the grounds that this outcome secures her a greater material benefit is not appropriate simply because this is not the mathematical function we expect her to act as if she were maximizing. The appropriate utility function is always the mathematical function a person acts as if she is trying to maximize, regardless of what she is actually trying to do or why.[22]

There are some critiques of utility theory that do not rest on simple misunderstandings, however. To understand these, it is first necessary to explain the two important assumptions of utility theory already mentioned: consistency and continuity. Consistency requires that if someone acts as if she prefers (on whatever grounds) A to B and B to C, then it follows she should also act as if she prefers A to C. While consistency seems reasonable, it has been shown that actual human beings do not always act as if they had consistent preferences. It is hard to imagine, however, that anyone would ever stick with genuinely inconsistent preferences once becoming aware of them. This is due to the problem of ‘improving oneself to death’. Suppose someone acts as if she prefers *A* to *B*. This being the case, she ought to be willing to pay a small sum less than their difference in value to exchange *A* for *B*. If, contrary to the consistency assumption, she also acts as if she prefers *B* to *C* and *C* to *A*, then by a long series of cycling exchanges it will be possible for her to spend herself into poverty, only to return to *A* in the end. Obviously, no one would be so foolish as to let this happen.[23]

Continuity is a somewhat more technical concept. The rough idea is as follows: Suppose someone acts as if she prefers *A* to *B* to *C*. Now suppose we offer her a choice between getting *B* for sure on the one hand, and a lottery yielding *A* with a probability *p* and *C* with a probability $1 - p$. Continuity requires that there exist a value for *p* such that she will act as if she is indifferent between getting *B* for sure and playing the lottery. (Note that continuity is compatible with being risk neutral, risk averse, or risk seeking.) It may be reasonable sometimes to have discontinuous preferences. One example would be lexical preferences: suppose a person believed liberty and efficiency are both good, but that liberty is infinitely more valuable than efficiency – or, to put it another way, that liberty and efficiency are ‘incommensurable’ goods. In this case, she would never be willing to trade off a gain in efficiency, no matter how great, for a loss in liberty, no matter how small. People often have, or at least profess to have, lexical preferences of this sort on a great variety of issues, though it is nearly impossible to determine the extent to which such preferences genuinely exist. (It may be that a person has never yet agreed to exchange a small loss of liberty for a large gain in efficiency, and indeed she may insist that she would never do so, but this in itself is not conclusive: who can say what a person would do in extreme situations?[24])

But let us concede that people at least some of the time engage in behavior violating the consistency and continuity assumptions of utility theory; the question, then, is the degree to which this presents a problem for RCT. Suppose that, in contrast with what I have been arguing, we regard game-theoretic rational choice models as attempts to provide intentional explanations of social phenomena. In this case, the existence of inconsistent or discontinuous preferences would indeed be troubling: it would suggest that at least some social phenomena cannot be shown to arise from the deliberate or intentional effort of human beings to maximize their utility. Moreover, any attempt to cordon off such cases as simply not amenable to rational choice analysis will seem to be an unjustified and arbitrary domain restriction on the rational choice approach.

I have argued that RCT should not be understood in this way, however. Rather, on the best available understanding, game-theoretic rational choice models should be seen as contributing to the construction of ordinary causal explanations. Consider again the Cox model. The social phenomenon to be explained is Duverger's Law, and the causes are shown to lie in the interaction of institutional structures with voters' beliefs regarding expected electoral results. Since the utility functions of the voters are assumed in any case, it would be strange to claim they were doing any explanatory work in the model. Cox's aim is not to defend any claim about voters' motivations, but rather to help unpack the causal processes leading from particular configurations of electoral law to the number of effective political parties. Understood in this way, it hardly matters whether the assumptions of utility theory are generally sound or not. So long as they roughly capture the expected behavior of a large number of voters, they serve well enough as tools for constructing a model of the causal mechanisms we are interested in better understanding.

Before moving on to critiques of the rationality assumption, let me note one other complaint sometimes directed towards utility theory: namely, that it treats utility functions as fixed givens, unexplained and not subject to revision. This is really a case of what (in a slight abuse of technical language) I refer to as the endogeneity problem, discussed in the Appendix.

II

A second group of complaints, often not clearly distinguished from the first, concerns not so much the construction of utility functions themselves, but rather the assumptions made with respect to what social actors *actually do* – regardless of the preferences they may or may not happen to have. The rationality assumption holds that, roughly speaking, discrete purposeful actors will optimize their utility functions, given whatever constraints they happen to face. What this entails depends on the situation, as was explained in part one of this article.

Critiques of the rationality assumption generally contend that human beings often do not, in fact, conform to the rules of rationality. Some of these critiques again result from confusion. For example, a common complaint is that rather than trying to maximize something, people actually engage in habitual, traditional, rule- or norm-bound, etc., behavior. To some extent, this may be true as a description of what people actually do.[25] But as we have seen, utility theory is agnostic with respect to what people are actually doing, so long as whatever it is can be represented *as if* it were an attempt to maximize some mathematical expression. In theory, there is no reason to believe rule-guided behavior cannot in general be so represented. For example, suppose someone always follows the rule, 'keep your promises, except when the other party defects first'. In this case, there are four outcomes to consider (see Figure 2). The numbers in each box represent a utility schedule over the possible outcomes for our promise-keeper. We may grant she is actually following the rule in question; nevertheless, it is the case that she acts *as if* she is trying to maximize the utility function specified, and this is all that utility theory requires.

Other critiques of rationality are more subtle, however, and do not rely on a misunderstanding of utility theory. Some of these include the following: First, that in many situations, the utility calculations required by rational choice models are far too complex for ordinary persons to actually carry out. Second, that even when people know what they want and can in principle carry out the necessary calculations, they do so erroneously. (For example, it is well-known that people systematically err in making certain probability calculations.) Third, that people at least some of the time act randomly, or contrary to what they themselves recognize to be their desired ends (as, for example, when caught up in the heat of passion, or suffering from weakness of will).[26] And lastly, that there are various 'paradoxes of rationality' – situations in which the principles of rationality seem necessarily to fail. For example: some desired outcomes cannot be achieved by trying to bring them about; some strategic games do not have equilibrium solutions;

	the other party:	
	Does not defect	Defects
Keep promise	1	0
Do not keep promise	0	1

Figure 2.

optimizing information-gathering can lead to an infinite regress; and so on.[27]

Now suppose we believe RCT attempts to provide intentional explanations of social phenomena. In this case, these problems would indeed be troubling. For example, many game theoretic models involve the use of extremely advanced mathematical techniques. Surely it is implausible to believe that actual social actors could (or would) correctly perform the calculations apparently required by the model. How then can any social phenomenon be explained with reference to intentional actions allegedly based on such calculations? It cannot. But this is only a problem so long as we persist in believing that RCT means to provide intentional explanations of those social phenomena. Once we see that the point of the model is only to unpack the causal mechanisms at work, it hardly matters: it is the structure of the situation in general that is being modeled, not rational decision-making in particular.

Similarly, consider the problem of emotions, passions, or weakness of will. Under the influence of these, a person might fail to perform the actions required to bring about outcomes she herself desires. Were RCT attempting to provide intentional explanations of social phenomena, this would indeed be a problem: outcomes would not have their origins in the intentional states of the social actors in question. But of course this is no problem for RCT when it is understood merely as a set of tools for constructing causal explanations of the usual sort. As I have argued, the utility functions in game-theoretic rational choice models perform little or no explanatory work; it hardly matters then whether we can draw an explanatory chain of the correct sort from the social phenomenon to be explained to the intentional states of particular social actors. We may simply assign social actors utility schedules that mimic whatever we expect them to do, without regard to whether those utility schedules actually reflect the desires and beliefs of the social actors in question.

From this point of view, the existence of rationality failure in certain cases does not matter much at all. By way of analogy, imagine that a medical researcher is testing the effectiveness of a new vaccination. Vaccination is causally connected with its medical benefits by virtue of how a healthy person's immune system normally reacts to the vaccine. The medical researcher's causal model, therefore, will include the assumption of a normal, healthy immune system, so that the effect of the vaccination itself can be isolated.

Surely it would be no objection to this research that not all people's immune systems function normally.[28] The role of the rationality assumption in constructing game-theoretic causal explanations of social phenomena is similarly limited, and the existence of rationality failures should similarly raise no objections to the model.

III

A third bundle of complaints directed against rational choice theory concern its supposed commitment to what is called *methodological individualism*. Earlier I suggested rational choice theory is in fact not so committed – that it only relies on a weaker *discrete purposeful actor* assumption. In order to assess this claim, however, we must first be clear about what, precisely, methodological individualism entails. Unfortunately, this is not so easy to figure out.

First, a clarification: no sensible person denies the existence of 'relational' or 'emergent' social phenomena. Let me explain. Suppose we have a bag of colored marbles, two-thirds of which are red. A distribution is obviously a relational property. That is to say, since an individual marble considered by itself can only be red or not red, the distribution of red marbles emerges only when a collection of marbles are considered together. At the same time, the distribution merely supervenes on the properties of the individuals: change the color of one of the individual marbles, and the distribution necessarily changes as well. Relational or emergent social phenomena of this sort often play a crucial role in rational choice models. Incomplete information models, for example, depend on the distribution of 'types' in a population of players; the Cox model discussed above employs the distribution of beliefs concerning likely electoral outcomes in determining results; and so on. Let us assume, then, that everyone concedes the existence and causal efficacy of relational or emergent social phenomena.

Now consider some large-scale social phenomenon social scientists might be interested in understanding – say, the agricultural production process in the USA (the example is not important). What might such an understanding look like? One point to be made here is that talk of 'the agricultural production process in the USA' must really be a sort of shorthand, for clearly 'the agricultural production process in the USA' is not a thing at all, in the way a bridge or a bird or a toaster is; rather, it is a vast conglomeration of actions, beliefs, dispositions, etc., of individual human beings. To be sure, this conglomeration happens to be patterned in a particular way, and thus counts as an emergent social phenomenon. Nevertheless, what we call 'the agricultural production process in the USA' is, strictly speaking, only a notional object supervening on an underlying pattern of actual, discrete social phenomena. A second point follows from the first: namely, that whatever the correct understanding of some large-scale social phenomenon happens to be, it must ultimately consist of a correct understanding of the individual actions, beliefs, dispositions, etc., of human beings constituting the large-scale social phenomenon in question. Of course, we are not always in a position to actually produce such an understanding, nor would it always be very useful, but this is neither here nor there: the issue is conceptual, not practical.

Often methodological individualism is taken simply to mean the assertion of these two interrelated points. In this form, methodological individualism can hardly be controversial, for indeed it is trivially true. Even Hegel accepts the doctrine in this form.[29] The advocates of methodological individualism may have something like this in mind, but their statements of the doctrine are not always perfectly clear. Consider the following two examples:

> We shall not have arrived at rock-bottom explanations of . . . large-scale [social] phenomena until we have deduced an account of them from statements about the dispositions, beliefs, resources and inter-relations of individuals. (Watkins 1959: 106)

> The elementary unit of social life is the individual human action. To explain social institutions and social change is to show how they arise as the result of the action and interaction of individuals. (Elster 1989: 13)

Now one might interpret these statements in the manner I have just suggested, and indeed both authors probably intend to assert something like the above-mentioned two points.[30] Unfortunately, this is not exactly what they say. To repeat, the two-fold assertion here is as follows: Considering some large-scale social phenomenon *X*, it is the case, first, that *X* is constituted by (or is shorthand for, or supervenes on) a patterned conglomeration of individual actions, beliefs, dispositions, etc., of human beings; and second, that understanding *X* ultimately means understanding these individual actions, beliefs, dispositions, etc., of the human beings in question.

Both authors above, however, appear to make an additional claim that the patterned conglomeration of individual actions, beliefs, dispositions, etc., of human beings constituting *X* must themselves ultimately be caused by (must arise from, must be deduced from) the actions, beliefs, dispositions, etc., of individual human beings. Sometimes methodological individualism is taken to be the assertion of this third claim (presumably in addition to the first two). On this alternate interpretation, however, the doctrine is not trivially true; quite the contrary, it must surely be false. Consider again our example of the agricultural production process in the USA. Now there are a great many reasons the patterned conglomeration of individual actions, beliefs, dispositions, etc., of human beings we call 'the agricultural production process in the USA' takes the particular shape and form it does. Many of these reasons are themselves the actions, beliefs, dispositions, etc., of individual human beings or, somewhat less directly, emergent social phenomena arising from the same. For example, the decision of a particular wheat farmer in Nebraska is shaped in part by the structure of her expectations regarding what other wheat farmers are likely to do. But also among these reasons are natural facts like the cycle of seasons. The cycle of seasons is neither directly nor indirectly a property of individual human beings, but clearly it affects (rather dramatically) the particular shape or form taken by the patterned conglomeration of individual actions, beliefs, dispositions, etc., of human beings we call 'the agricultural production process in the USA'. Therefore, a complete explanation of this large-scale social phenomenon would *not* have 'arrived at rock-bottom', so to speak, unless or until it included relevant natural facts of this sort. Accordingly, we must reject this claim, and with it the second interpretation of methodological individualism.

If RCT is supposed to be controversially committed to something called 'methodological individualism', this commitment must be interpreted in some other way, for neither of the above interpretations seem worthy of controversy (for opposite reasons). Consider next a much weaker version of the third claim, asserting not that *all* the causes of large-scale social phenomena must, directly or indirectly, be the actions, beliefs, dispositions, etc., of individual human beings, but rather only that *some* of them must be. This much weaker version of claim number three is, in my view, almost certainly correct, though unlike the first two claims it is not *trivially* so. On the contrary, it is apparently denied by 'pure structuralists' who hold that complete explanations for large-scale social phenomena are still possible after dropping out any references to individual human beings. Most rational choice theorists, quite sensibly, would endorse methodological individualism in this form.[31] But they need not. The discrete purposeful actor assumption is weaker even than this weak version of claim three: it requires only that purposeful actors of some sort – be they human beings, firms, political parties, or something else – have at least some causally efficacious role in explaining large-scale social phenomena. Now perhaps we have reasons to doubt whether anything other than an individual human being could satisfy the conditions of utility theory, and thus behave as a discrete purposeful actor in the required sense: Arrow's theorem, for example, tells us that no (non-dictatorial) decision rule can guarantee consistent collective preferences. It does not follow, however, that consistent collective preferences do not exist. Just how common they are is an empirical question, and nothing in RCT entails denying their existence all together.

Before concluding this discussion, I should perhaps consider a less rigorous version of the complaint: rational choice theorists, one might complain, are more skeptical than they ought to be about the possibilities of cooperative or collective action, while at the same time being overly naive concerning the structural constraints social actors must contend with. As a result, rational choice theorists tend to minimize the causal role structures play in determining social outcomes. Quite apart from the fact that nothing in rational choice theory requires the alleged over- and under-emphases, this charge is in any case unfounded. On the contrary, typical game-theoretic rational choice models are in a sense *all about* structure. To use the example of Cox's model once more, surely electoral rules, the distribution of voter beliefs, and so on, count as structure if anything does, and the whole point of his model was to explain why different configurations of these structures predictably generate different outcomes.

One final note: the problem of methodological individualism is sometimes confused with the separate and more general problem of endogeneity. Thus it is occasionally argued that rational choice theory renders structures exogenous to its models, when they ought to be endogenous. This objection is considered in the Appendix.

IV

Finally, a number of complaints have been directed not so much at the formal assumptions made by rational choice theorists, but rather at the practical execution of what is sometimes called (misleadingly, in my view) the 'rational choice project'. These complaints are, roughly, that RCT falls short of the generally accepted standards for good scientific research. The most commonly discussed issues include: post hoc theory development, unfalsifiability, and arbitrary domain restriction. Each of these can be disposed of in short order.[32]

Some people are of the opinion that empirical social science research ought to be problem-driven rather than theory-driven: in other words, rather than seeking out problems to analyze with one's favorite theory, researchers ought to begin by thinking about what the real problems are, and then seek out the best theoretical tools for analyzing them. It is also sometimes thought desirable that researchers produce bold and unexpected predictive claims, rather than merely provide *ex post* accounts for known past events. Rational choice theory, it is argued, violates both of these methodological desiderata: to a large extent, the research conducted by rational choice theorists consists merely of providing after-the-fact accounts of how known social phenomena might have arisen from the intentional pursuit of self-interest by social actors. Such work is theory driven rather than problem driven, and it does not yield bold or unexpected predictive claims. This, roughly, is the 'post-hoc theory development' objection.

I do not intend to take a position, one way or the other, on such views about the appropriate way of doing social science; certainly, many people emphatically reject the picture just described. But I will point out that as an objection to RCT, the objection is misguided. The mistake consists in thinking of rational choice as an explanatory theory at all – a mistake stemming from the view that the goal of RCT is to provide intentional explanations of social phenomena. Cox does not argue that rational choice theory explains Duverger's Law, but rather that by using the tools of RCT we can construct a model showing how electoral laws and voter expectations jointly explain Duverger's Law. The assumptions of RCT are neither confirmed nor rejected by his research, nor indeed is this the point. Only so long as we persist in mistakenly believing this *is* the point will it appear that such work is necessarily theory driven and mere after-the-fact story telling.

The second set of methodological objections involves the problem of forming conclusive tests of RCT. For a theory to be testable, it must have clearly-defined criteria for falsification. The utility theory component of rational choice theory violates this requirement, however: as defined, utility is an unobservable phenomenon and hence utility functions are unfalsifiable. Whenever a social actor fails to maximize her utility function, utility theory states merely that we have assigned her the incorrect function; the correct function is the one we in fact expect her to try to maximize. This being the case, nothing, it would seem, could falsify utility theory nor, by extension, the rational choice models employing it.

Were rational choice theorists purporting to provide intentional explanations of social phenomena, this would indeed be a problem. Since utility functions are simply defined merely as those functions we expect social actors to act as if they are attempting to maximize, an intentional explanation of social phenomena grounded on utility maximization would indeed be (almost) perfectly circular. But this is not the role utility functions play in well-constructed game-theoretic rational choice models. They merely summarize in a concise mathematical expression our expectations regarding their behavior insofar as it is relevant for the social phenomenon we are interested in understanding. The empirical content of the model, as I have argued, lies in comparative statics: it is the structure of the model itself being tested, not the specification of utility functions.

The final set of methodological objections considered here concern the problem of what is called 'arbitrary domain restriction'. Curiously, this objection flatly contradicts the second. It holds that RCT has in fact been falsified in a number of contexts, the most famous (already mentioned) example being its alleged failure to explain voter turnout. Rather than admit defeat, however, rational choice theorists arbitrarily restrict the domain of social phenomena subject to rational choice explanation to save the theory. This, for some reason, is thought to be objectionable.

This complaint is confused on many levels. To begin with, if RCT is unfalsifiable (as the second objection suggested), it cannot have been falsified. And indeed, RCT as such is very nearly unfalsifiable; utility theory is designed precisely to ensure this. But in any case, as has already been made clear, RCT is not supposed to be a testable theory at all: rather, it is a set of conceptual tools for constructing testable causal models. The fact that RCT itself cannot be falsified should therefore be regarded a strength, and not a weakness. Even granting, however, that there are interesting problems not easily amenable to rational choice analysis, it is far from clear why this should be viewed as an objection to the approach as a whole. Consider a parallel: Certainly, not all interesting social phenomena are easily quantifiable. It follows that the statistical methods of multiple regression will not be particularly helpful in explaining such phenomena. No one would refer to this as an ‘arbitrary domain restriction’ on multiple regression, or view it as an objection to such methods as such. Multiple regression is simply a bundle of tools, useful when useful, not useful when not useful. I have argued throughout this article that rational choice theory should be viewed in exactly the same way – as a set of methodological tools, particularly useful for constructing causal explanations of many interesting social phenomena. We have no good reason to believe RCT will necessarily be useful in explaining *all* interesting social phenomena, nor should it matter if it is not.

Conclusion

RCT is not a unified, monolithic, universal theory of social phenomena. Rather, it is a set of tools that sometimes help social scientists in their efforts to understand and explain social phenomena. The analogy with the statistical techniques of multiple regression, just mentioned, is worth repeating: No one talks of a ‘multiple regression project’ as if multiple regression were some broad theory about how the social world is put together, something that could be confirmed or rejected by evidence. Nor does the fact that these tools are not always useful warrant our throwing them out.

Granted, there might be over-enthusiastic practitioners of statistics, and also of RCT, who persist in employing their favored techniques even when they are less helpful. What is worse, there might even be some who claim that social phenomena which cannot be studied using their favored techniques must be uninteresting, unimportant, not a valid subject of social science research, or even non-existent. (I write ‘might be’ because the claims in question seem so obviously false, I am not persuaded anyone would ever assert them seriously.) The fault then lies with the practitioners, not with the tools they employ.

One final comment. I do not want to suggest that the mountain of existing critical work on RCT is entirely worthless. One could imagine a theory – let us call it ‘rational choice theory*’ – that substituted the doctrine of psychological egoism or something else for the assumptions of utility theory. Social scientists might then attempt to employ rational choice theory* in constructing intentional explanations of some social phenomena. At least some of the time they might succeed. The critical work on RCT is valuable in showing us just how limited we should expect such successes to be.[33] Of course, it does not follow that intentional explanations as such are problematic. For any given social phenomenon, there may be none, one, or more than one good explanation. When there are good explanations, the best one may be intentional, and sometimes an intentional explanation from psychological egoism specifically. It is extremely unlikely, however, that there would ever be one and only one best form of explanation for all possible social phenomena (and certainly not intentional explanations from psychological egoism).

My goal in this article has been to argue that if one looks carefully at the best research actually conducted within the loosely-defined rational choice paradigm, one will find that, properly understood, it does not provide or even purport to provide intentional explanations of social phenomena. Rather, RCT is employed merely as a set of tools useful in developing straightforward causal explanations of social phenomena. So understood, much of the literature on either side of the rational choice controversy is simply besides the point.

APPENDIX

Statistical models sometimes run into a difficulty referred to as the *endogeneity problem*. This problem arises when an independent or explanatory variable turns out in fact to be partly caused by the dependent variable it was supposed to explain. For example, it has proved difficult to show that campaign spending in US Congressional races improves a candidate's chances of winning. A candidate who expects to win easily will often spend less than a candidate who believes she needs to spend a lot in order to win; thus the level of spending (the alleged independent variable) is partly determined by the probability of a candidate's winning (the dependent variable).

There is a lesson to be drawn from this problem, as follows: it is impossible to explain everything simultaneously. All explanations therefore involve explaining the less familiar with reference to the more familiar. To put it another way, not everything can be endogenous to a model: on the contrary, for any model to have explanatory force, some things (independent or explanatory variables in particular) must necessarily be exogenous. Generally speaking, whatever a model tries to explain will be endogenous with respect to the model, and everything else will be exogenous. The line between the endogenous and the exogenous is thus always relative to what one is interested in explaining. This being the case, it cannot by itself be an objection to some model in particular that it treats some things exogenously. (Of course, if a person claimed her model explains *X* when in fact *X* is exogenous to her model, that would be an objection. But who would be so foolish as to make such an elementary mistake?)

Now consider two objections sometimes made against game-theoretic rational choice models. On the one hand, it is complained that such models treat utility functions as fixed givens, unexplained and not subject to revision. On the other hand, it is complained that such models leave social structures unexplained by merely defining them as 'rules of the game'. These are both complaints about where the line has been drawn between the endogenous and the exogenous. Roughly, both objections state that whereas a model treats *X* as exogenous, it ought to treat it as endogenous.

Since the complaint cannot be against exogeneity as such, we must try harder to give sense to these objections. I can imagine two plausible interpretations: First, it might be objected that a model treats *X* as exogenous in explaining *Y*, but that it is really *X* and not *Y* that is interesting and deserving of explanation. Second, it might be objected that a particular type of model (in this case, rational choice models) cannot possibly explain *X*, and so must necessarily treat *X* as exogenous.

Now I am far from convinced that the social phenomena typically in question here are necessarily beyond the scope of RCT. Preferences and social structures are themselves social phenomena. It is likely that the processes giving rise to such social phenomena involve at least some constrained purposeful action on the part of discrete social actors. To the extent that this is the case, there may presumably exist game-theoretic rational choice models helpful for understanding the processes in question. At any rate, I see no reason for denying such a possibility a priori.

That said, it does not particularly matter whether my intuition on this point is correct or not. Suppose there is some set of social phenomena *X* whose members are *necessarily* not amenable to rational choice analysis. Why should this matter? RCT would still be useful for analyzing social phenomena not in *X*. Only if the members of *X* are seen as particular interesting, and the social phenomena not in *X* as particularly uninteresting, would this be a real problem. (Even then, the problem would only be that RCT is not useful for understanding what happens to interest us, and not that it is invalid as such.) Fortunately, this does not seem to be the case. There are clearly cases in which the constrained purposeful action of discrete social actors gives rise to social phenomena we are interested in. RCT is useful for understanding these, and that should be enough.

NOTES

1. I leave aside here non-empirical explanations, as for example a normative explanation for why some state of affairs is just or unjust. Other candidates are either reducible to one or a combination of these three basic modes (historiography), or else are incomplete parts of an explanation (hermeneutics).
2. Causal mechanisms are of course often presumed, without being stated, on the basis of a strong correlation between variables alone, but this is neither here nor there for our purposes. A strong correlation itself is generally supposed to be good evidence for some unspecified (and perhaps unknown) causal mechanism; indeed, this is precisely why we regard the correlation as interesting in the first place.
3. Roughly, functional explanations are successful when a causal loop in some iterated series of events can be identified. Typically, the causal loop will be a selection mechanism of some sort (e.g., natural selection in evolutionary biology). For discussion, see Merton (1957), Stinchcombe (1968), Elster (1979: 28–35, 88–103), Elster (1983a: ch. 2), and Elster (1989: chs 8–9).
4. This would happen if it could be further shown that first-order intentional states themselves always arise deterministically, and not, say, from voluntary second-order intentional states. (An example of the latter might be a voluntary desire to cultivate in oneself a desire for classical music.)
5. See Searle (2001: chs 2–3), for discussion and arguments to the effect that intentional states are not causally sufficient antecedent conditions of human action.
6. This view is ubiquitous to the point where citation is unnecessary, but some representative examples from diverse literatures include: Elster (1986: 12–22), Pettit (1993: 264–82), Green and Shapiro (1994: 20–3), and Turner (2003: ch. 20). Satz and Ferejohn (1994) describe the intentional explanation view as the 'received interpretation' of RCT, though they go on to criticize it. Most often, writers to not explicitly state their view, but their arguments vis-a-vis the rational choice debate clearly imply the intentional explanation view.
7. There are of course exceptions to this generalization: in particular, see Satz and Ferejohn (1994) and Diermeier (1995). This article owes much to these authors.
8. As it happens, I believe that functional rational choice explanations are seriously problematic, but for reasons unconnected with the main argument here.
9. Cf. Elster (2000: 24–5), who identifies three similar core assumptions, though they are interpreted differently in important respects. By contrast, Green and

Shapiro's (1994: 14–17) discussion of five core assumptions is thoroughly confused.

10. In other words, the decision to do *A* need not actually be due to a preference for the expected outcome of doing *A*: this is merely how it is described by a utility function. This important point will be further elaborated below.
11. Named after John Nash (1951) because in a legendary paper he demonstrated that under a very general set of conditions, the existence of at least one such equilibrium point is guaranteed. For a fairly compact reconstruction, see Gibbons (1992: 29–48) or Binmore (1992: 319–29).
12. Other auxiliary principles might include 'sub-game perfection', 'Bayesian updating', etc. On this extended understanding of rationality, see Elster (1986: 4–12, 2000: 28–31).
13. Herein lies my principle disagreement with MacDonald (2003).
14. Here we assume chance does not intervene, and outcomes can be read off deterministically once the players have chosen their respective strategies. There are, however, easy techniques for modeling chance which need not detain us here.
15. The mathematical details involved in constructing utility functions were worked out in von Neumann and Morgenstern (1943). For a nice restatement, see Luce and Raiffa (1957: ch. 2).
16. Some believe that game theory predicts *no one* will vote, but this is not correct. Zero turnout cannot be a Nash equilibrium, because if no one else votes, then each would want to vote because her vote will be decisive. The correct game theoretic solution involves what are called mixed strategies: see Morrow (1994: 212–16).
17. Let me note that I follow Elster's (1989: esp. ch. 1) account of causal explanation here: he argues that an event is causally explained by citing another earlier event, together with a causal mechanism connecting them. Others follow Hempel (1965), who argues that an event is explained when it is the deductive consequence of general laws plus initial conditions. It is unfortunately beyond the scope of this paper to elaborate on this dispute.
18. See especially Segal and Spaeth (2002). I am grateful to an anonymous *Rationality and Society* reviewer for suggesting this example.
19. And, I would go on to add, with good reason, for we do not generally have access to the relevant data – namely, the prior intentional states of actors.
20. The term 'psychological egoism' derives from Sidgwick (1907: 39–42), though his terminology is somewhat different. See Feinberg (1993) for a discussion. Note that a different set of problems results when 'thick rationality', rather than psychological egoism, is substituted for orthodox utility theory. Objections to thick rationality as a theory of human motivation include the problems of framing, ideological bias, preference manipulation, and so on. See Elster (1983b).
21. Not all other-regarding motivations are altruistic: for example, revenge can be other-regarding, especially when it harms the revenging agent as well as the victim. Expressive and symbolic action is not necessarily non-instrumental: for example, I might participate in a political rally not because I enjoy doing so intrinsically, but in order to express solidarity.
22. Naturally, it need not be the same mathematical expression for all persons. Some are under the curiously mistaken notion that RCT requires an assumption that all people are motivated by the same things. On the contrary, it is rare to find rational choice models that make such an assumption in practice.

23. A slightly more complicated case arises due to inconsistency over time: see Elster (1979: 65–77).
24. For discussion, see Elster (1979: 124–7) and Posner (2000: ch. 11).
25. Particularly persuasive as a description of what people actually do is the notion of 'satisficing': often, rather than seeking a globally optimal solution, social actors tend to look around until they find a merely satisfactory solution, and settle on that. See especially Simon (1982).
26. See discussions in Elster (1979: 139–41, 157–79, 1989: chs 5, 7) and Searle (2001: ch. 7).
27. See Elster (1979: 117–23, 133–7, 150–3, 1983b: ch. 2).
28. Having established the general effectiveness of the vaccine, our medical researcher might go on to do a second study with the goal of figuring out how close to normal a person's immune system has to be for the vaccine to be effective, and how many people fall within that range. This second study *would* be analogous to testing an intentionality model of social behavior. My contention is that rational choice theory models are generally more like the first sort of study than the second.
29. In the introduction to his *Philosophy of History*, Hegel argues that the path of history must proceed through the needs, passions, characters, and talents of individual human beings.
30. Elster, at least, regards methodological individualism as 'trivially true' (1989: 13), which suggests he adheres to this interpretation of the doctrine.
31. The reasons for accepting this form of methodological individualism, and thus rejecting pure structuralism, are nicely discussed in Whitmeyer (1994). Roughly speaking, the argument runs as follows: Consider any allegedly pure structuralist explanation for a large-scale social phenomenon. Now suppose that human psychology were wildly different than we typically think it is: would this really have no effect on our explanation? Surely it would. But then any complete explanation for a large-scale social phenomenon must include references to individual human beings. I am grateful to a *Rationality and Society* reviewer for pointing out this reference and its relevance for the discussion here.
32. This discussion in this section largely follows the objections raised by Green and Shapiro (1994), and the generally sound replies given by Diermeier (1995).
33. In this genre, the best work by far is that of Jon Elster; I will not dispute any of his critiques of rationality so far as they go. My only disagreement with Elster lies in his regarding rational choice models as intentional explanations of social phenomena.

REFERENCES

Binmore, K. 1992. *Fun and Games: A Text on Game Theory*. Lexington, MA: D.C. Heath and Company.

Cox, G.W. 1997. *Making Votes Count: Strategic Coordination in the World's Electoral Systems*. New York: Cambridge University Press.

Diermeier, D. 1995. 'Rational Choice and the Role of Theory in Political Science.' *Critical Review* 9: 59–70.

Dworkin, R. 1986. *Law's Empire*. Cambridge, MA: Belknap Press.

Elster, J. 1979. *Ulysses and the Sirens: Studies in Rationality and Irrationality*. New York: Cambridge University Press.

Elster, J. 1983a. *Explaining Technical Change: A Case Study in the Philosophy of Science*. New York: Cambridge University Press.

Elster, J. 1983b. *Sour Grapes: Studies in the Subversion of Rationality*. New York: Cambridge University Press.

Elster, J. 1986. 'Introduction.' In *Rational Choice*, ed. Jon Elster, pp. 1–33. New York: New York University Press.

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

Elster, J. 2000. 'Rationality, Economy, and Society.' In *The Cambridge Companion to Weber*, ed. Stephen Turner, pp. 21–41. New York: Cambridge University Press.

Feinberg, J. 1993. 'Psychological Egoism.' In *Reason and Responsibility: Readings in Some Basic Problems of Philosophy*, ed. Joel Feinberg, pp. 461–72. Belmont, CA: Wadsworth Publishing.

Gibbons, R. 1992. *Game Theory for Applied Economists*. Princeton, NJ: Princeton University Press.

Green, D.P. and I. Shapiro. 1994. *Pathologies of Rational Choice Theory: A Critique of Applications in Political Science*. New Haven, CT: Yale University Press.

Hempel, C. 1965. *Apects of Scientific Explanation*. New York: The Free Press.

Luce, R.D. and H. Raiffa. 1957. *Games and Decisions: Introduction and Critical Survey*. New York: Wiley.

MacDonald, P.K. 2003. 'Useful Fiction or Miracle Maker: The Competing Epistemological Foundations of Rational Choice Theory.' *American Political Science Review* 97: 551–65.

Merton, R. 1957. *Social Theory and Social Structure*. New York: The Free Press.

Morrow, James D. 1994. *Game Theory for Political Scientists*. Princeton, NJ: Princeton University Press.

Nash, J.F. 1951. 'Non-Cooperative Games.' *Annals of Mathematics* 54: 286–95.

Pettit, P. 1993. *The Common Mind: An Essay on Psychology, Society, and Politics*. New York: Oxford University Press.

Posner, E.A. 2000. *Law and Social Norms*. Cambridge, MA: Harvard University Press.

Satz, D. and J. Ferejohn. 1994. 'Rational Choice and Social Theory.' *Journal of Philosophy* 91: 71–87.

Searle, J.R. 2001. *Rationality in Action*. Cambridge, MA: MIT Press.

Segal, J.A. and H.J. Spaeth. 2002. *The Supreme Court and the Attitudinal Model Revisited*. New York: Cambridge University Press.

Sidgwick, H. [1907] 1981. *The Methods of Ethics*. Indianapolis, IN: Hackett Publishing.

Simon, H. 1982. *Models of Bounded Rationality*, vol. 2. Cambridge, MA: MIT Press.

Stinchcombe, Arthur L. 1968. *Constructing Social Theories*. New York: Harcourt Brace & World.

Turner, Jonathan H. 2003. *The Structure of Sociological Theory*, 7th ed. Belmont, CA: Wadsworth/Thomson Learning.

von Neumann, J. and O. Morgenstern. [1943] 1953. *The Theory of Games and Economic Behavior*, 3rd ed. New York: John Wiley.

Watkins, J.W.N. 1959. 'Historical Explanations in the Social Sciences.' *British Journal for the Philosophy of Science* 8: 104–17.

Whitmeyer, J.M. 1994. 'Why Actor Models are Integral to Structural Analysis.' *Sociological Theory* 12: 153–65.

FRANK LOVETT is an assistant professor of political science at Washington University in St Louis. He would like to thank Paul MacDonald and Joseph Parent for many discussions concerning the article's themes; the participants of the May 2003 Social Theory Convention, Tampa Bay, for their comments; and the anonymous *Rationality and Society* reviewers for their engaging criticism and suggestions. He would also like to express a special debt to those from whom he learned rational choice theory, especially Nolan McCarty and Charles Cameron. His research interests include theories of justice, civic republicanism, judicial politics and the rule of law, and social science epistemology.

ADDRESS: Department of Political Science, One Brookings Drive, St. Louis, MO 63130, USA [email: flovett@artsci.wustl.edu]