Explanation, Formal Models, and Rational Choice*

Scott Ashworth†

June 21, 2023

In *Theory and Credibility,* Chris Berry, Ethan Bueno de Mesquita, and I argued that formal modeling and causal inference are natural complements, pointing to a variety of ways researchers can bring them closer together. We knew this might be a hard sell, since theorists and empiricists are often suspicious of the other's enterprise:

> While, in principle, nearly everyone agrees that theory and empirics ought to work together, in practice, each side feels the other isn't holding up its end of the bargain. On the one hand, a group of theoretically minded scholars is baffled and dismayed by the empirical turn towards research designs for credibly answering narrow causal questions. Why, they wonder, are empiricists obsessed with carefully answering uninteresting questions, rather than doing work that speaks to theoretical questions? On the other hand, a group of empirically minded scholars is similarly baffled and dismayed by theorists' focus on abstract models built on, from their perspective, demonstrably false assumptions. Of

*A previous version was prepared for the Conceptualizing and Contextualizing Formal Theory Conference. I received helpful comments and questions from many participants at the conference, including Ethan Bueno de Mesquita, Bobby Gulloti, Itai Sher, Branislov Slantchev, and my discussant, Scott Tyson. I also thank Harvey Lederman, Kevin Zollman, and the organizers for comments on the subsequent draft, and Mark Schroeder for pointing me toward some helpful references. The paper draws extensively from my book with Ethan Bueno de Mesquita and Chris Berry, *Theory and Credibility.* They have not reviewed this paper, and are not responsible for any of its flaws.

†Harris School of Public Policy, University of Chicago. sashwort@uchicago.edu

> what use, they wonder, can such models be for explaining the world or guiding empirical inquiry? [p. 3]

I stand by the book's conciliatory tone and message, but I'm under no illusion that the rift will be fully closed anytime soon. My aim here is to explain why. I'll do so by diagnosing one cause of the rift, explicating the issues that are woven into it, and suggesting that different stances toward those issues will resist easy resolution.

To provide some focus for the discussion, I'll reconsider the quarter-century-old debate surrounding *Pathologies of Rational Choice Theory.* Green and Shapiro (1994) argued that rational choice theory had yet to deliver on its promise to improve the empirical study of politics. The book sparked a flurry of responses from scholars standing up to defend rational choice theory (Friedman, 1996). Fully addressing all of the back-and-forth is too much a single paper, so I'll focus on just one thread: Green and Shapiro's original book, the response by Ferejohn and Satz (1996), and Green and Shapiro's (1996)'s reply to Ferejohn and Satz's argument.

The debate is rooted in a disagreement about how to think about formal rational choice models in political science. Implicit in Green and Shapiro's argument is a conceptualization something like this:

> Models are tools to generate possible causal explanations, which should be tested against behavioral data.

Ferejohn and Satz don't reject this, but they do think it is incomplete. They explicitly defend a conception of models like this:

> Models are **a unified collection of** tools to generate possible **intentional and** causal explanations, which should be tested against behavioral data.

I suggest that contention over the bolded additions remains an undercurrent in exchanges between quantitative and formal political scientists.

No one disputes the importance of those additions in the practice of formal rational choice theory. Green and Shapiro (1994, p. 20) emphasize the role of intentional explanations: "rational choice explanations are typically formulated by reference to individual intentions". They also at least gesture in the direction of the desiderata of a unified collection of models. They accuse rational choice theory of a commitment to universality, a commitment that "results from its proponents' conception of scientific advance, which is thought to occur when generalizable results can be shown to follow from analytic propositions derived from axioms" [p. 24].

The dispute is about the value of those additions. In Section 4 I will return to the debate over *Pathologies* and discuss the textual basis for my characterization of the disagreement. But first, I need to step back and clarify some points. Section 1 discusses two notions of explanation in political science: causal and intentional. Section 2 draws on Giere's (1988; 2006) model-based view of scientific theorizing to elucidate how unified collections of formal models support causal explanation. Section 3 examines the rational choice strand of formal modeling to assess how much it is genuinely oriented towards intentional explanation.

Finally, in Section 5 I argue that these different conceptions ground different attitudes towards models with mixed empirical support. A researcher interested in intentional explanation may choose to work on improving an empirically flawed account even though an empirically adequate, non-intentional account is already on the table. In this way, formal rational choice models' orientation toward intentional explanation is a persistent source of friction between quantitative and formal political scientists.

1 Two Kinds of Explanation

Social scientists try to explain social phenomena. We want to go beyond saying what does happen, to shed light on *why* it happens. But that simple why-question hides two different

conceptions—one oriented around causes and the other oriented around intentions.[1]

Both conceptions can be motivated by objections to the covering law account of explanation. Przeworski and Teune's (1970) influential textbook can stand as our example of how social scientists embraced that account. On page 19, they write:

> For example, why does Monsieur Rouget, age 24, blond hair, brown eyes, a worker in a large factory, vote Communist? To explain the vote of M. Rouget, one must rely upon general probabilistic statements that are relevant for voting behavior and have been sufficiently confirmed against various sets of evidence. The particular features of M. Rouget must be used as the first premise of the explanation:
>
> > M. Rouget is a worker and
> > works in a large factory and
> > is young (24 years old).
>
> The second premise consists of a conjunction of general statements describing with a high likelihood the behavior of skilled workers, employees of large factories, and young persons. (No interaction is assumed.)
>
> > One out of every two workers votes Communist; and employees of large organizations vote Communist more often than employees of small organizations; and young people vote communist more often than older people.
>
> Therefore, it is likely that
>
> > M. Rouget votes Communist.[2]

[1] This is not to say that intentional explanations are not causal explanations—they are. (The locus classicus of the causal theory of intentional action is Davidson (1963); Paul (2010) is an important recent defense of the theory.) What is crucial for the argument is that there are plenty of causal explanations that are not intentional and, as will be clear below, many of the explanations offered by political scientists are causal but not intentional.

[2] This is an instance of what Hempel (1962) calls inductive-statistical explanation.

Don't laugh—we still do work like this all the time. This is, after all, just a regression tree, a key component of important predictive tools from contemporary machine learning (James et al., 2021). Moreover, I think that much of the most valuable empirical political science being done right now is devoted to carefully establishing basic descriptive facts about political life (e.g., Little and Meng, 2023; Westwood et al., 2022). But few contemporary social scientists would say this work is explanatory.

There are two different objections to taking a raw statistical generalization as an explanation. One points to innovations in quantitative empirical work; the other points to a central aspect of the dominant tradition in formal theory. I'll take them up in that order.

1.1 Causal Explanation

The first objection: covering-law explanations do not properly account for causation. Przeworski and Teune's positivist tradition treats explanation and prediction as symmetric. One and the same argument is a theoretical prediction if the event mentioned in the conclusion is future and is an explanation if the event mentioned in the conclusion is in the past. But there is no symmetry thesis for causation and prediction. Were M. Rouget admitted to his local ICU, we could predict poor health two weeks later. But the explanation of his later poor health would not be his stay in the ICU—if anything, that stay presumably improved his health. Whatever health event caused him to enter the ICU also caused his poor health. It is such causal facts that the first objector wants in an explanation.

From this point of view, what a student of political behavior wants to know in Przeworski and Teune's example is whether or not taking a factory job has a positive causal effect on voting Communist. Answering such a question requires making certain kinds of counterfactual comparisons. Would M. Rouget continue to vote Communist if he switched to an office job? What if he had taken an office job straight out of school? You can't an-

swer these questions just using data on M. Rouget, since there is only one of him.[3] But the decades since Przeworski and Teune wrote, enormous progress has been made on estimating averages of such causal effects. Social scientists now often use randomized experiments to answer such questions, and are sensitive to the costs and benefits of a variety of features of observational studies that might warrant causal inference when experiments are not possible.

Once a causal notion of explanation has taken hold, it's natural to seek deeper explanations in terms of intermediate causes, often called *mediators* (Imai, Keele and Tingley, 2010; Green, Ha and Bullock, 2010). Perhaps a fuller story of M. Rouget goes like this. The Communists actively canvass around factories. Thus taking the factory job caused M. Rouget to be canvassed, and, it turns out, that canvasing activity caused him to vote for the Communists. Had he taken an office job, he would not have encountered any Communist canvasing, and would have voted for the Socialists. In that case, the causal effect of the factory job on voting Communist is positive, and that effect is mediated by party canvasing activity. The fundamental problem of causal inference still applies, so a quantitative political scientist is never going to learn those casual facts about M. Rouget in particular. But there is still hope for learning population-level facts about such networks of causal facts.

1.2 Intentional Explanation

The second objection: covering-law explanations do not properly account for intentions. Przeworski and Teune's explanation ignores that M. Rouget *made a decision* to vote for the communists. The statistical generalization fails to do this fact justice. We want to know more: What does M. Rouget believe about the Communist platform? What does he think about the ability of the party's candidate? Does he worry about wasting his vote if he decides instead to vote for a smaller left-wing party? And so on. We want, in other words,

[3]But see Halpern (2016) for an approach that can sometimes answer such questions using an auxiliary structural causal model a la Pearl (2009).

to understand M. Rouget's reasons for voting Communist.

An *intentional* explanation is one that gives the actor's reason for acting as they do. Here, a reason is an appropriate constellation of beliefs and desires that show why the act makes sense for the agent. Appropriate means two things at least. One is that the act is a good way to achieve the desire given the belief. The other is that *these* beliefs and desires actually lead to the act. That is, they don't just happen to be appropriate to an act undertaken for a different reason.

This is not to say those beliefs and desires have to enter the conscious awareness of the actor. What is crucial for an intentional state to actually lead to the act is that the intentional state is part of a true causal explanation of the act. At least two considerations animate rational-choice theorists to seek an intentional explanation rather than some other kind of causal explanation.

The first consideration is that intentional explanations ground a rich set of counterfactual claims. For example, when asked if he considered neoclassical choice-theoretic foundations essential in economic modeling, Robert Lucas (1999, p. 159) responded:

> No. It depends on the purposes you want the model to serve.... [I]f one wants to know how behavior is likely to change under some change in a policy, it is necessary to model the way people make choices. If you see me driving north on Clark Street, you will have good (though not perfect) predictive success by guessing that I will still be going North on the same street a few minutes later. But if you want to predict how I will respond if Clark Street is closed off, you have to have some idea of where I am going and what my alternative routes are—of the nature of my decision problem.

Lucas's point is that different reasons for driving north on Clark ground different constellations of causal explanations of how he would respond to various obstructions.

The intentional explanation, then, gets the level of detail right. Imagine subjecting poor

Lucas to a battery of experiments. See where he turns if Clark is closed at Belmont, and at Roscoe, and so on. In principle, we could build up to the complete story of how closures of Clark cause him to take different side streets. But that story would miss the real pattern in his (actual and counterfactual) behavior—he wants to get to Wrigley Field to see the White Sox play the Cubs, and he's choosing a sensible route to get there. Dennett (1989) argues that examples like Lucas's are ubiquitous, making the intentional level of analysis ideal for understanding behavior quite broadly.

The other consideration goes further, making intentional explanations valuable independent of any connection to prediction in fine-grained counterfactuals. Dray (1957, p. 128, emphasis in original) formulates the point this way: "Only by putting yourself in the agent's position can you *understand* why he did what he did." This is not practical advice about imagining yourself into the shoes of the agent to figure things out. Rather, Dray is formulating "certain *conditions which must be satisfied* before a historian is prepared to say: 'Now I have the explanation.'" A little bit later, he expands on the point by distinguishing two points of view from which we can approach an action. One is "as a spectator", looking at it as part of a pattern or regularity. He contrasts this with adopting the "standpoint of an agent" [p. 140].

What Dray is saying here, I think, is that predicting behavior is not the only way we relate to one another. People are interested in interpreting, and empathizing with, one another. Suppose Alice quits her job, and Bob asks why. What he wants to know is her reason for quitting—how, from her point of view, quitting makes sense. We don't leave these interests behind when we start doing social science. And we should not let anyone's methodological pronouncements prod us into suppressing them. Interpretation and empathy involve construing people intentionally. And thus, social science has a permanent place for intentional explanations.[4]

[4]A tradition, starting perhaps with Dilthey and running through such contemporary political theorists as MacIntyre and Taylor, takes this last set of considerations to the limit, going so far as to argue that causal

2 Formal Models

Suppose you're curious why a party canvasser can persuade the workers at M. Rouget's factory to vote for the Communists. Your next-door neighbor is a theorist who studies communication, so you ask her. Being bad at pedagogy, she launches into formalism: "Consider the collection $\langle \Theta, M, A, u^S, u^R, \pi \rangle$, where Θ is a set of states ...". Could that possibly be the beginning of a helpful answer?

Some social scientists build formal models because they (we) believe the answer is yes. Here's the idea. The theorist identifies some target of inquiry in the world, or *target* for short. The theorist then builds a mathematical structure—a *model*—that is somehow similar to the target. The model (e.g., the sender-receiver game partially described above) *represents* the target (e.g., communication between the canvasser and a worker). The hope is that studying that mathematical structure is an indirect but useful way to learn about the world.[5]

These models are typically not sui generis, but are built as part of theoretical traditions. Those traditions embody principles that constrain and guide the modeling enterprise. Giere (2006, p. 62), writing about fundamental principles in physics, characterizes these principles "as general templates for the construction of more specific models... to the principles, one adds what I'm here calling 'specific conditions,' the result being a more specific, but still abstract, object. The principles thus help both to shape and also to constrain the structure of these more specific models." We see a theoretical tradition at work in the first paragraph of this section. The theorist is modeling communication between the scientist and the legislator as a Bayesian game (Myerson, 1991). The definition of a game creates one of Giere's general templates. The theorist specifies a game by filling in blanks corresponding to things like players (S and R), possible actions (the sets M and A), types (Θ), beliefs (π), and payoffs (u^S and u^R). Building models according to the principles helps us see the

explanation is inapt for human action.

[5](Godfrey-Smith, 2006) is a good overview of this strategy in scientific theorizing.

world through a particular theoretical lens.

Whatever theoretical lens we're looking through, the analytical process has two steps. The first step is inquiry into the model—analyzing the model without reference to a particular target. The second step is inquiry with the model—relating the model to the world for the purpose of explanation and assessment.[6]

To inquire into the model, a theorist asks questions about the model itself. At this stage, she has no concern for model's relationship to any particular target. She simply tries to figure out, using some combination of analytical and computational tools, implications of the model's assumptions. In doing so, the theorist would like to be able also to give some explanation of the results, to say why the model's assumptions have those implications. She wants to show how the various parts of the model fit together and interact to produce the result.[7]

To figure this out, she considers nearby models—models that are alike in many respects but differ in others—in an attempt to identify the *crucial features* of the model. A couple of common theoretical practices help the theorist find these crucial features. First, models usually contain components that are well understood in the relevant tradition. Second, the theorist works incrementally. She changes one component of the model at a time, using established models as benchmarks. Both practices support discovery of perspicuous accounts of why the model's results are as they are.

To inquire with the model, by contrast, the researcher does have to concern herself with the model's similarity to some target. I won't attempt a full analysis of what similarity amounts to, but instead will rest content with clarifying the scope of the claimed relationship.[8] A model should be similar to the target in a very particular way. Certain actors, situations, and relationships in the model must meaningfully represent analogous

[6]Here I am paraphrasing Morgan (2012, p. 31).

[7]This is what theorists typically have in mind when the ask for the mechanism or story of a model.

[8]Frigg and Nguyen's (2020) encyclopedia article is an excellent introduction to the debates about interpreting similarity in model-based science.

actors, situations, and relationships in the target. Call those the *representational features* of the model. Not everything in the model is meant to be representational. Any theoretical model has *auxiliary features* that are not descriptive of the target, but help keep the model tractable. In a sender-receiver game, for example, the assumption that the sender's payoff function is constant in the message would be representational, capturing a sender who faces no consequence for, and feels no qualms about, lying. The common assumptions of uniformly distributed states and quadratic payoff functions, on the other hand, would be auxiliary.

Inquiry into a model yields implications about relationships between objects in the model. Some of those implications depend on the representational features, but are robust to changes in the auxiliary features. Those implications are suitable for assessing similarity of the model and the target. But other implications depend crucially on an auxiliary feature of the model. There is no reason to expect to see the implication reflected in the world.

Some similarities between target and model are effectively forced by the theoretical tradition. A theorist trying to explain the effect of canvassing on vote choice probably could not convince a reasonable interlocutor that the worker was similar to the sender and the canvasser to the receiver. But other similarities are more open. The states, for instance, might correspond to possible facts about the party platforms, to candidates' experience, or even to the time of day the polls open.

My example and any knowledge you have of my own work probably combine to make you read "model" above as "analytically solved rational choice model". But the basic view I'm outlining applies more broadly. What I've said so far also covers, mutatis mutandis, traditions such as computational modeling, evolutionary approaches, and mathematical schemas for formalizing causal relationships, such as those of Pearl (2009) and Imbens and Rubin (2015).[9] All of these approaches support empirical work into causal explanations in

[9]Blair et al. (2019) treat such schemas as theoretical models.

the same way: The theorist explores the space of possible relationships inside the model by varying its crucial features. To the extent that these features are representational, the model's bundle of possible relationships should be similar to the counterfactual structure of the target.

3 Rational Choice Models

The dominant tradition in formal political theory is rational choice theory. That theoretical tradition demands models be built with the elements of intentional explanation—agents and their opportunities, beliefs, and desires. So formal political theory at least gives the appearance of offering intentional explanations. How seriously should we take that appearance?

Before taking up that question, I want to be very clear about one point. I am not claiming that an intentional model must be a canonical rational choice model. Far from it. Kahneman and Tversky's (1979) prospect theory, Simon's (1956) bounded rationality, and Sen's (2004) reasoned choice are all alternative theoretical traditions organized around the elements of intentional explanation. I suspect that the account I'm about to give extends to them as well, but I leave the work of cashing out that suspicion to the future.

Back to the main plot. Suppose a theorist has used noncooperative game theory to model some political phenomenon. That model specifies the players' opportunities, preferences, and beliefs about the state of the world. Inquiry into the model typically looks for an equilibrium, i.e., a collection of strategies each of which is optimal in light of the other. A strategy bundles together facts about the player's plan of action and the other players' beliefs about that plan. By nailing down the crucial features of the model, the theorist identifies intentional explanations inside the world of the model.

There is a range of views about how seriously we should take these intentional states in the model.

The *strict revealed preference view* treats the preferences and beliefs instrumentally. As Binmore (2008, p. 19, italics in original) puts it:

> In revealed preference theory, it isn't true that Pandora chooses b rather than a *because* the utility of b exceeds the utility of a. This is the Causal Utility Fallacy. It isn't even true that Pandora chooses b rather than a because she prefers b to a. On the contrary, it is because Pandora chooses b rather than a that we say that Pandora prefers b to a, and assign b a larger utility.

On this view, preferences and beliefs are modeling devices that perspicuously capture facts about behavior. That behavior is representational, but the preferences and beliefs are not.

Many theorists, though, disagree with this construal of revealed preference theory. Virtually all will go part way with Binmore, agreeing that his Causal Utility Fallacy really is a fallacy. But the elimination of preferences in favor of choice behavior is more controversial.

First, a clarification. Binmore writes "it is because Pandora chooses b rather than a that we say Pandora prefers b to a". This is ambiguous. One way to read it is as an epistemic claim: The best evidence we have about preferences is choice. I suspect most theorists would agree with Binmore's claim on that reading. But another way to read it is as a semantic claim: The meaning of 'prefers b to a' is 'chooses b over a'. If these two expressions really have the same meaning, then we can freely substitute one for the other without changing the truth of the larger statement. Read that way, the semantic claim is controversial among formal rational-choice theorists.

An objection to the semantic claim arises from the way preferences enter rational choice explanation. After acknowledging that the strict revealed preference view is something of the official view, Blume and Easley (2008, p. 886) write:

> Much of economics involves invisible hand explanations; aggregate market behavior emerges from the decisions of many agents. Whether the invisible hand lifts the cup aloft or knocks it over, economic explanation entails explaining

> how it coordinates for good or ill the motives and interests of diverse individual actors. These kinds of questions call for explanations based on the motivations of economic actors, which purely behavioralist explanations cannot provide. So economists in practice take an intentional view.

To expand on the penultimate sentence: It makes sense to say that "Pandora chooses b over a because she prefers b to a". That appears to be an explanation, telling us why Pandora chooses b. But if the semantic claim is correct, then we can substitute "chooses b over a" for "prefers b to a". That substitution yields "Pandora chooses b over a because she chooses b over a". That is in no way an explanation of Pandora's choice. Blume and Easley maintain that explaining choices in terms of motivations is essential to the promise of formal rational choice models. So, while paying lip service to the strict revealed preference view, formal theorists must be taking an intentional view.

Binmore bites the bullet on this point:

> The price of abandoning psychology for revealed-preference theory is therefore high. We have to give up any pretension to be offering a causal explanation of Pandora's choice behavior in favor of an account that is merely a description of the choice behavior of someone who chooses consistently. [p. 20]

For this response to work, the appeal to consistency must be available prior to any consideration of intentions.

Sen (1993) argues that such an appeal is not in fact available: making sense of the consistency conditions already involves non-choice factors like motivations. The strict revealed preference theorist cashes out references to preferences being transitive as an oblique way of talking about contraction consistency: if y is chosen from $\{x, y, z\}$, then y is also chosen from $\{x, y\}$.[10] But just what is inconsistent if x is chosen over y from $\{x, y\}$? Remem-

[10] In the text I'm restricting attention to preference relations without indifference. In the general case, contraction consistency would be replaced with the weak axiom of revealed preference.

ber, the answer must not appeal to anything beyond choice behavior! Sen argues that this challenge in insuperable:

> Statements A and not-A are contradictory in a way that choosing y from $\{x, y, z\}$ and x from $\{x, y\}$ cannot be. ... Given some ideas as to what the person is trying to do ..., we might be able to "interpret" these actions as implied statements. But we cannot do that without invoking such an external reference. There is no thing as purely internal consistency of choice. [p. 499]

We can see Sen's point by considering Binmore's attempt to dismiss one of Sen's examples. Discussing an example in which the alternative z involves cocaine, he writes:

> The reason that she violates [contraction consistency] without our finding her behavior unreasonable is that snorting cocaine isn't an *irrelevant* alternative for her. ... If we want to apply the theory of revealed preference to her behavior, we must therefore find a way to formulate her decision problem in which no such hidden relationships link the actions available in her feasible set A with either her beliefs concerning the states in the set B or the consequences in the set C. [p. 11]

The intentionalist can agree with Binmore's advice while wondering what right he has to give it. The option to snort cocaine is relevant to the decision maker because of her constellation of intentional states—her beliefs about just which strangers would propose such an opportunity and her preferences about hanging out with them.[11]

The upshot of this discussion is simple. Formal rational choice theorists rarely live up to the dictates of the strict revealed preference view. Instead, they take the intentional parts of their models seriously, as candidates for being representational features. And when that works, when the agents in the model are similar to people in the target, then inquiry with

[11] Thoma (2021) argues that this step of individuating the options is the only concession a defender of the strict revealed preference view need make to mentalism.

the model provides understanding—the reasons people in the world act as they do will be similar to the reasons the actors in the model act as they do.

Thus formal rational choice modeling has two distinctive characteristics. First, it is more obviously indirect than most political science, studying not the target directly but a made-up mathematical structure that is similar to the target.[12] Second, it is committed to intentional explanations. This commitment might well involve trade-offs with causal explanation. I turn now to some suggestive evidence that these two differences conspire to create misunderstanding between quantitative and formal political scientists.

4 A Pathological Debate?

Green and Shapiro examine rational choice models and associated empirical work in four substantive domains—voter turnout, collective action, voting in legislatures, and platform choice in elections. Their assessment on all four cases in negative: "despite its enormous and growing prestige in the discipline, rational choice theory has yet to deliver on its promise to advance the empirical study of politics" [p. 7]. They argue that these failures have a methodological origin: "the weaknesses of rational choice scholarship are rooted in the characteristic aspiration of rational choice theorists to come up with a universal theory of politics" [p. 6].

Green and Shapiro support this diagnosis with an extended discussion of rational choice explanation. They point out that rational choice models provide intentional explanations, while implicitly rejecting the claim that intentional explanations are in any way privileged among possible explanations of human behavior. This rejection is most clear in their discussion of Riker's defense of rational choice theory on intentionalist grounds:

> The argument with which Riker confronts the Skinnerian behaviorist can be

[12] Debates about the external validity of survey responses or populations recruited on MTurk indicate that a lot of empirical political science might be pretty indirect as well.

> pressed equally against the rational choice theorist's appeal to the primacy of intentions and preferences. What determines them! Perhaps they are products of chemical reactions in the brain or of cultures or of institutional orders. To say that intentions and choices are the building blocks from which explanations of human behavior should be fashioned rests on nothing more than a conjecture that they are in fact the basic determinants [p. 186].

In their response to Green and Shapiro, Ferejohn and Satz (1996, p. 74) resist this glib treatment of intentional explanation:

> Social-science explanations must, we claim, be compatible with intentional descriptions of human agents. That is to say, it is a necessary constraint on a social science explanation that the agents whose actions are required to make it work be capable of forming the intentions appropriate to those actions. A successful social-science explanation presupposes that the behavior in question be describable as being intentionally brought about by human agents seeking goals and holding beliefs.

Their argument for this position echo the discussion above of Lucas and Dray:

> We claim, first, that in everyday life, human agents must and do make use of intentionalist interpretation in order to make attributions about mental states. Second, we claim that this form of intentionalism must satisfy a pragmatic test—it must allow its holders to make good predictions as to how others will behave in a wide variety of settings—so that it is "generative," and cannot consist of mere tautologies. In order to satisfy these requirements, we claim that this "folk intentionalism" must be describable in universalistic terms. Finally, we claim that successful intentional scientific accounts must "track" folk intentionalism, and therefore inherit its universalistic features [p. 79].

Disappointingly, Green and Shapiro's (1996) reply to their critics engages with this argument only in a dismissive endnote, pointing to the gap between a bare intentional explanation and an explanation in terms of one of the standard models of rational choice. But we can get some insight into how they might respond by considering the other prong of their critique.

According to Green and Shapiro, the universalistic aspirations of rational choice theory "results from its proponents' conception of scientific advance, which is thought to occur when generalizable results can be shown to follow from analytic propositions derived from axioms" [p. 24].

It's important to be clear about what's right and what's wrong in this claim. It's true that the model-building principles of formal rational choice theory are often studied axiomatically. But there would be no hope of deriving, say, Krehbiel's (2010) pivotal politics from some general axioms of rational choice politics. The better reading of universalism here is a Giereian one. Rational choice theory is a family of formal models, built according to principles that are often presented axiomatically. Why study that family of models, built in the framework of those principles? The rational choice theorist would respond with some combination of its track record of producing successful causal explanations and its ability to provide intentional understanding. But when Green and Shapiro (1994, p. 194) raise this question, they say this:

> An even weaker version of the family-of-theories argument is the claim that rational choice theory is not a theory at all; rather, it is an "approach," a "methodology" or a "paradigm," and as such it cannot be tested. If this argument is taken seriously, the question becomes, Why choose the rational choice approach rather than a different one? The answer, presumably, must rest on an appeal to the predictive success of the hypotheses that the approach yields.

Green and Shapiro seem to value intentional explanation only heuristically. And they

are surely not alone in that. The failure of Green and Shapiro and Ferejohn and Satz to have a meeting of the minds, I want to suggest, reveals a persistent fault-line in disciplinary discussions of rational choice models. Many (most?) theorists are willing to sacrifice some empirical adequacy to get better understanding. Many (most?) empiricists are not.

All of this said, Green and Shapiro land some body blows against the empirical work they discuss. And a reasoned willingness to trade off empirical adequacy for understanding cannot go so far as to neglect the empirical side altogether. I'll conclude with some brief remarks on this.

5 Conclusion

Let's take stock. We want to explain behavior in political settings. Distinguish two dimensions of fit between a proposed explanation and the target explanandum. One dimension concerns whether an explanatory theory represents only in terms of behaviors—e.g., votes cast, treaties signed, rebellions joined—or also in terms of the agents' reasons for those behaviors. The other dimension concerns how much the explanation tracks (factual and counterfactual) variation in behavior in the target. How should we weight these dimensions?

The graphs in Figure 1 and 2 plot the performance of two theoretical accounts of some phenomenon. The horizontal axis measures well the account supports an intentional interpretation of the phenomenon. The vertical axis measures how well the causal structure of the account is reflected in data about behavior. An ideal theoretical account would perform well on both dimensions. There are few, if any, models in political science that currently live up to that ideal. We face trade-offs when assessing explanations.

The two gray lines reflect different tradeoffs scholars might be willing to make between the two dimensions. The solid line represents my construal of Green and Shapiro's position. Intentional interpretability is valued only instrumentally, just to the extent that it leads

to better behavioral adequacy. The dashed line represents a position that treats both dimensions as independently valuable. Some degree of behavioral adequacy can be sacrificed for improvements in intentional interpretability.

This second view is a moderate version of Ferejohn and Satz's view. Recall that they write: "Social-science explanations must, we claim, be compatible with intentional descriptions of human agents". This is a strong claim, arguably inconsistent with the constant tradeoff represented by the dashed line. That constant tradeoff implies that absent any intentional interpretability at all, there is a compensating improvement in behavioral adequacy that makes the trade worthwhile. In what follows, I will ignore this wrinkle and treat the dashed lines as Ferejohn and Satz's view.

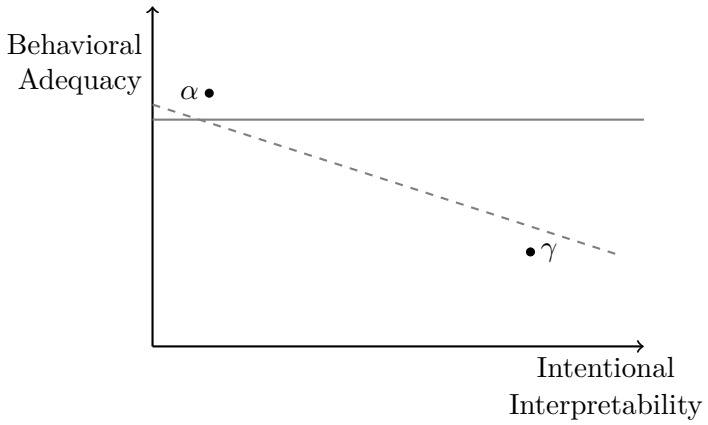

Figure 1

Figure 1 schematically represents the situation as Green and Shapiro saw it in 1994. The rational choice theory, γ, scores well on intentional interpretability, but abysmally on behavioral adequacy. The traditional political science theory, α, lacks intentional interpretability. But its behavioral adequacy is much better than the rational choice alternative. As drawn, the rational choice theory's behavioral adequacy is so compromised that Ferejohn and Satz should agree that α is a better theoretical account than γ.

In spite of this agreement about γ and α, there can easily be disagreement about what

to do next. Green and Shapiro are inclined to forget about γ as a non-starter of an account. But Ferejohn and Satz could reasonably continue working on theoretical accounts related to γ. They would be looking for an alternative more like β in Figure 2.

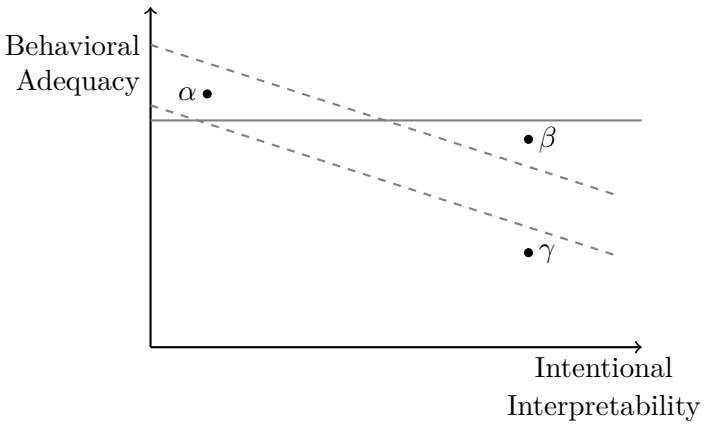

Figure 2

One can imagine Green and Shapiro frustrated by this. We already have an account with better behavioral adequacy than γ. An improvement to β won't dethrone α on that dimension. Why then devote so much effort to the pursuit of β? Even if that effort pays off, α will be the best account. Perhaps the moderate Ferejohn and Satz are just bewitched by the shiny methods of formal theory.

Ferejohn and Satz reject that diagnosis. They see a situation with two flawed theoretical accounts. Lack of behavioral adequacy is indeed a problem for γ. But an account that resists intentional interpretability as much as α does is also a problem.

Moreover, they see γ as a reasonable starting point for doing better. Because it of the way it is embedded in a Giereian tradition of intentional modeling, the theorist has a solid understanding of how the components of the model fit together to produce its results. And she can work incrementally, treating γ as a baseline to understand the effects of modifying those components one by one. These considerations do not guarantee that starting from γ will succeed, and it is probably too much to hope for that doing so will succeed quickly.

But if it does succeed, Ferejohn and Satz want to suggest, the wait will have been worth it.

Readers can judge for themselves which side of this debate is validated by the literatures we discuss in Ashworth, Berry and Bueno de Mesquita (2021). My point now is not to indulge in hindsight, but to highlight a division that we still live with.

References

Ashworth, Scott, Christopher R. Berry and Ethan Bueno de Mesquita. 2021. *Theory and Credibility: Integrating Theoretical and Empirical Social Science*. Princeton: Princeton University Press.

Binmore, Ken. 2008. *Rational Decisions*. Princeton: Princeton University Press.

Blair, Graeme, Jasper Cooper, Alexander Coppock and Macartan Humphreys. 2019. "Declaring and Diagnosing Research Designs." *American Political Science Review* 113(3):838–859.

Blume, Lawrence E. and David Easley. 2008. Rationality. In *The New Palgrave Dictionary of Economics*, ed. Steven N. Durlauf and Lawrence E. Blume. London: Palgrave Macmillan pp. 884–893.

Davidson, Donald. 1963. "Actions, Reasons, and Causes." *Journal of Philosophy* 60(23):685–700.

Dennett, Daniel C. 1989. *The Intentional Stance*. Cambridge: MIT press.

Dray, William H. 1957. *Laws and Explanation in History*. New York: Oxford University Press.

Ferejohn, John and Debra Satz. 1996. Unification, Universalism, and Rational Choice Theory. In *The Rational Choice Controversy: Economic Models of Politics Reconsidered*, ed. Jeffrey Friedman. New Haven: Yale University Press pp. 71–84.

Friedman, Jeffrey. 1996. *The Rational Choice Controversy: Economic Models of Politics Reconsidered*. New Haven: Yale University Press.

Frigg, Roman and James Nguyen. 2020. Scientific Representation. In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Spring 2020 ed. Metaphysics Research Lab, Stanford University.

Giere, Ronald N. 1988. *Explaining Science: A Cognitive Approach*. Chicago: University of Chicago Press.

Giere, Ronald N. 2006. *Scientific Perspectivism*. Chicago: University of Chicago Press.

Godfrey-Smith, Peter. 2006. "The Strategy of Model-Based Science." *Biology and Philosophy* 21(5):725–740.

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

Green, Donald and Ian Shapiro. 1996. Pathologies Revisited: Reflections on Our Critics. In *The Rational Choice Controversy: Economic Models of Politics Reconsidered*, ed. Jeffrey Friedman. New Haven: Yale University Press pp. 235–276.

Green, Donald P., Shang E. Ha and John G. Bullock. 2010. "Enough Already About "Black Box" Experiments: Studying Mediation Is More Difficult than Most Scholars Suppose." *The Annals of the American Academy of Political and Social Science* 628(1):200–208.

Halpern, Joseph Y. 2016. *Actual Causality*. Cambridge: MIT Press.

Hempel, Carl G. 1962. Deductive-Nomological vs. Statistical Explanation. In *Scientific Explanation, Space & Time, (Minnesota Studies in the Philosophy of Science, vol. III)*, ed. Herbert Feigl and Gordon Maxwell. Minneapolis: University of Minnesota Press pp. 98–169.

Imai, Kosuke, Luke Keele and Dustin Tingley. 2010. "A General Approach to Causal Mediation Analysis." *Psychological methods* 15(4):309.

Imbens, Guido W. and Donald B. Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. New York: Cambridge University Press.

James, Gareth, Daniela Witten, Trevor Hastie and Robert Tibshirani. 2021. *An Introduction to Statistical Learning*. New York: Springer.

Kahneman, Daniel and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica* 47(2):263–291.

Krehbiel, Keith. 2010. *Pivotal Politics: A Theory of U.S. Lawmaking*. Chicago: University of Chicago Press.

Little, Andrew T. and Anne Meng. 2023. "Subjective and Objective Measurement of Democratic Backsliding.".

Lucas, Jr., Robert E. 1999. Interview with Brian Snowden and Howard R. Vane. In *Conversations with Leading Economists: Interpreting Modern Macroeconomics*, ed. Brian Snowden and Howard R. Vane. Northhampton MA: Edward Elgar pp. 145–165.

Morgan, Mary. 2012. *The World in the Model: How Economists Work and Think*. New York: Cambridge University Press.

Myerson, Roger B. 1991. *Game Theory: Analysis of Conflict*. Cambridge, MA: Harvard University Press.

Paul, Sarah K. 2010. "Deviant Formal Causation." *Journal of Ethics & Social Philosophy* 5(3):1–24.

Pearl, Judea. 2009. *Causality*. Second ed. New York: Cambridge University Press.

Przeworski, Adam and Henry Teune. 1970. *The Logic of Comparative Social Inquiry*. New York: Wiley Interscience.

Sen, Amartya. 1993. "Internal Consistency of Choice." *Econometrica* 61(3):495–521.

Sen, Amartya. 2004. "Incompleteness and Reasoned Choice." *Synthese* 140(1/2):43–59.

Simon, Herbert A. 1956. "Rational Choice and the Structure of the Environment." *Psychological Review* 63(2):129–138.

Thoma, Johanna. 2021. "In Defence of Revealed Preference Theory." *Economics & Philosophy* 37(2):163–187.

Westwood, Sean J., Justin Grimmer, Matthew Tyler and Clayton Nall. 2022. "Current Research Overstates American Support for Political Violence." *Proceedings of the National Academy of Sciences* 119(12).