Behavioral Political Economy, Argumentation, and Democratic Theory

DIMITRI LANDA

Abstract

A developing research program of behavioral political economy can help shed light on important social and political practices that fall outside the strict rational actor model but that are of central importance to democratic theory. Those practices include the deliberative activities of argumentation, information acquisition, and learning. Game theoretic models and experimental studies of collective decisions that are part of the behavioral political economy tradition offer insights into the strategic implications of these practices, linking them to ideological polarization and measures of the informational quality of individual and collective choices. In so doing, they help generate comprehensive assessments of these practices and their institutional influences, thus buttressing the normative philosophical arguments.

Keywords: behavioral political economy, argumentation, deliberative democratic theory

Scholars who identify themselves as political philosophers tend not to write in math, while those who identify themselves as political economists often do.1 With that distinction come important sociological differences, but one would
be hard-pressed to find principled differences that would not run into either logical or empirical difficulties: at the very least, this is true of the old distinctions between the preoccupations with questions of ought vs. questions of is, or between the inquiries into the nature of concepts vs. explanations of behavior. But the sociological differences are real, and they are certainly related to the fact that the languages in which these fields operate are not easily mutually accessible. Thus, explicit efforts to build bridges between them are valuable when there are important substantive connections to be had. I will argue that such connections exist between some of the recent developments in behavioral political economy (henceforth, BPE) and some of the important preoccupations of democratic, and especially deliberative democratic, theory, and while the relationship between BPE and political philosophy is almost entirely unexplored, there is considerable value in pursuing it.

The points of contact and productive engagement between political economy and political philosophy have extended to a variety of theoretical developments, from the earlier work on social contract and repeated games, distributive justice and fair division, and Arrovian social choice to the more recent work on the Condorcet Jury Theorem, doctrinal paradox/discursive dilemma, and deliberative democratic theory. Almost all of those engagements have relied on the traditional political economy approaches, developing considerable insight into a number of distinct social and behavioral phenomena. But there are important classes of such phenomena that involve behavior central to political philosophy concerns yet are difficult to approach within the traditional political economy models. I have in mind, in particular, behavior related to individual and group choices involving the deliberative activity of argumentation. Such behavior takes place in strategic settings and, arguably, reflects purposeful choices by individuals responding to the distinct incentives implicit in the underlying political institutions. As such, it is a natural focus for political economy models, and the BPE contributions I discuss seek to account for it by marrying the methodological strengths of game-theoretic modeling to more descriptive behavioral micro-foundations.

One of the inspirations for BPE is behavioral economics, which has taken off as one of the most significant field-defining developments in economics in the last one-and-a-half decades. Behavioral economics is predicated on the recognition of the importance of departures from the classic assumptions of rational agency for explaining a variety of behavioral phenomena identified by empirical scholars. At its best, it proceeds by bridging the micro-level laboratory findings of experimental economists with the aggregate behavioral patterns found in observational
studies. The model of theoretical inquiry as micro-macro bridging, is, of course, the gold standard of economic reasoning and has been the gold standard in political economy since its inception as a field. In this sense, theoretical BPE is of a piece with that tradition, both in the form of inquiry and in the expected rigor, but whereas micro is assumed to be the classic form of rational agency in classic political economy, BPE substitutes for that assumption a model (or models) of agency with elements explicitly derived from experimental evidence.

The issues that have been central to the developments in BPE concern learning and belief updating. Laboratory evidence suggests that even for the relatively well-educated, beliefs in response to new information change in ways that are significantly different from the Bayesian ideal entailed in the classic rational actor model. These micro-level findings are of particular relevance to deliberative democratic theory, for which much hinges on how citizens respond to each other’s arguments. Explicitly modeling these micro-level differences in learning can be shown to lead to political choices that are distinct from what one may expect of rational agents, and, set in the context of strategic political relationships with other citizens, can shed light on distinct classes of political phenomena that are of special interest to democratic theory.

II

There are multiple senses in which the rationality assumptions entailed in the classical model of rational agency are fairly minimal. I will mention two of them that I think have had the most traction with political economists. The first is that the bottom line with respect to both preferences and beliefs is consistency, and who doesn’t want to be consistent? As Dagfinn Føllesdal notes in a well-known formulation: “Man has rationality as a norm, as a second-order disposition of the following kind: once one becomes aware that one has fallen into irrationality, one will tend to adjust one’s belief, attitudes, and actions such as to make them more rational. . . . Man is a rational animal in the sense that man has rationality as a norm.” This formulation of rational agency is attractive in its minimalism, and it can be used to construct philosophical arguments for the more stringent rational actor assumptions in rational actor models: in repeated social settings,
people may get to rationality through experimentation and learning; if we want stable predictions, strict rational-agent assumptions may be precisely what we want.\(^5\)

The other sense of minimalism of the classic rational actor model is that it furnishes a kind of natural parsimonious baseline that may be wrong in fact, but provides a generalizable point of comparison. Ken Binmore provides the following statement of this rationale: “No discussion of the organization of human societies is possible without taking a view on what people are like. Any such view is bound to be wrong because it is impossible to encompass the richness and diversity of human nature in a few sentences of text. Nevertheless, if any serious analysis is to be attempted, one cannot follow the example of authors like Elster who seek to list all the multifarious factors that influence the way collective decisions are made. One needs to commit oneself to a specific view about which of the various influences that act upon us are really crucial to understanding what makes us tick.”\(^6\) The view that he has in mind and that is shared by many social scientists, likely owing to the recognition of the normative force of rationality, is the classic rational actor model.

These two arguments are quite forceful, and in many settings and with respect to a long list of questions of interest to political economists, I think they do point to the classical rational actor model as the desirable model of agency. Further, the analytical tools of game-theoretic reasoning encapsulate the idea of consistency in a strategic setting and, in that, are indispensable to contemporary political economy. As I argued elsewhere, those tools are crucial for understanding the consequences of different democratic institutions and developing a coherent normative theory of institutional choice; if political philosophy seeks to be able to make normative claims on behalf of some institutions and against others—claims that go beyond the obvious—than an engagement with political-economic analysis is essential.\(^7\) Moving from the general and hardly controversial claims that institutions should maximize social welfare or select policies that best fit the aggregate information available in society to arguments implicating particular institutions requires developing expectations about how individuals will respond to each other under the incentives those institutions create. One of the meta-lessons of modern political economy is that those responses often fall short of normatively convenient and socially optimal; the first-best institutions are often not implementable. Normative endorsements of institutions that do not engage with an equilibrium analysis of what is possible and stable are, thus, likely to be irrelevant beyond purely ideal-theoretic constructions.
Rather than rehearsing these arguments here, I will take their conclusion at face value. Instead, I will argue that there are specific phenomena that are of central importance to political philosophy that fall away from the formal analysis if one proceeds with the classical rational agent model. If one wants to understand those phenomena, one needs to be willing to amend that model, but it can be, and, for the reasons indicated, is best done without giving up game-theoretic tools. The idea of a game-theoretic equilibrium as a fixed point of the (perceived) best responses is defined for a given set of beliefs, even where those beliefs may depart from the efficiently updated responses to information. In other words, a behavioral political-economic analysis that focuses on the analysis of the implications of those beliefs does not need to surrender the idea of the normative status of consistency or the analytical tools of modern political economy that are critical for normative institutional analysis.

As for the values of parsimony and generalizability: they are fundamentally pragmatic, and if the point of the argument is that a given model is not useful for understanding a certain class of phenomena of interest, the value of understanding the phenomena must dictate the tools. A somewhat later statement by Binmore suggests a similar sensibility: “rational decision theory is only a useful positive tool when the conditions are favorable. Economists sometimes manage to convince themselves that the theory always applies to everything, but such enthusiasm succeeds only in providing ammunition for skeptics looking for an excuse to junk the theory altogether.” The key intuition of BPE is that the methodological choice is not between the rational actor model and baroque theory, and that the red line at the classical rational actor model does not separate a beautiful edifice from pandemonium. In this case, one can have what one wants, while living within one’s means.

So, what is to be gained from the particular interventions of BPE, and what does that do for political philosophy? I will focus on two sets of phenomena that I think are fundamental for democratic theory and that we understand better precisely because of the BPE’s analysis of those phenomena within the methodological setting just discussed, that is, specific micro-grounded departures from the classical rational agent model, set in the context of strategic interactions under particular institutional conditions, and evaluated using game-theoretic tools. The first set of such phenomena encompasses deliberative
behavior that is effectively “black-boxed” or assumed away by the classic rational agency assumptions. The second concerns particular deliberative practices that may persist because they do not run into corrections or challenges to inconsistency that may otherwise be expected to rationalize them. Consider them in turn.

The best and most widely accepted interpretations of the optimal choice assumptions is that they are summaries of unmodeled processes. Some describe them in terms of selection pressures that generate behavior that looks as if it were explicitly chosen to be optimal, but that may simply be the behavior that survives better than the behavior that is suboptimal. Others suggest that the processes may be more internalist—agents responding to arguments from others or following the examples of others when those examples appear to lead to higher welfare. The rational agent assumptions summarize the behavior that such processes arguably culminate in—under favorable circumstances—but they leave those processes themselves outside the model.

Actions that instantiate those processes are, however, of particular interest to democratic theory, and in particular, to deliberative democratic theory. In terms of informational contexts, theorists of deliberation envision deliberative interactions as a fairly encompassing category. But a particularly prominent ideal of citizen deliberations is one of a succession of arguments seeking to persuade “on the merits”—by “the force of a better argument.” Such a persuasion may entail a revelation of previously unknown facts, but a more resonant description is, probably, one of arguments seeking to expose explicit or implicit biases, by identifying unwarranted leaps of judgment or inference. Classical rational agency assumptions imply that individuals are not subject to such leaps and biases; all conclusions that may be derived from the accepted arguments or positions are already accepted and held, and those that cannot be so derived are not accepted. In other words, deliberation as many theorists of deliberation—among them, John Rawls who viewed the reasoning in the original position in those terms, but also Joshua Cohen, Bernard Manin, and others—would describe it, cannot, without contortion, be apprehended within the strict classical rational-actor model.

The second set of phenomena is closely related to the first but also to the behavioral implementation of the ideal of consistency expressed in Follesdal’s argument. Consistency and perceptions of consistency are not one and the same, and the rhetorical force of insisting on consistency begins to dissipate once we realize how easy it is to perceive consistency, or even more, how easy it is to convince oneself that there is consistency when it is not there. When the informational feedback received by the agents is reinforcing, rather than
counteracting, the potentially inefficient beliefs, agents may lack the mechanisms necessary to perceive inconsistency. A salient way in which this may occur is when agents are controlling the information to which they are exposed and systematically choose to restrict that exposure in a way that makes correction toward more efficient learning difficult and reinforcement of prior biases easier. When that happens, the bite of the “rationality is the norm” argument is clearly weaker, and the strict rational agent assumption becomes more awkward. Mechanisms of this kind, however, look exactly like what critics of selective biased information exposure suggest stand behind political polarization—a phenomenon of singular political prominence in present-day American politics and one of central importance to democratic theory. If the deliberative account of polarization is right, that would suggest that rather than reinforcing mutual understanding and toleration, as deliberative theorists hope, social mechanisms of deliberation may bring about an altogether opposite effect.

If the deliberative account of polarization is right, that would suggest that rather than reinforcing mutual understanding social mechanisms of deliberation may bring about an altogether opposite effect.

If correction to the rational may be undermined by systematically biased information acquisition, it may also be strained by individuals’ enjoyment of higher utility under inefficient learning. While the latter possibility may sound paradoxical, it need not be so, as we will see in a moment, in a strategic setting. Both of these possibilities, then—one, arguably, decreasing individual and social utility, while the other increasing it—reinforce the view that deliberation as “argumentation on the merits” may have considerable durability in the face of the rationality-policing mechanisms.

III

What, if anything, can political economy do to shed light on these phenomena pushing the boundaries of the rational actor model? Quite a bit—under the broad rubric of BPE. The goal of such an inquiry is to construct micro-models that can give rise to these phenomena as a proof of concept and then to derive their behavioral and institutional implications in strategic settings.

Departures from the classical rational actor model that sustain the nontrivial exercise of persuasion “on the merits” may take a number of distinct paths. But the more attractive will be those that, while behaviorally
plausible, remain broadly committed to rationality as a behavioral norm and incur minimal costs to parsimony and generalizability. The theory of *deliberation as self-discovery* developed over the last decade follows such a path. It is built on a model of agency in which agents learn from arguments that draw logical inferences from propositions they already hold true. However, they learn only when directly persuaded, that is, when the speaker’s claims “resonate” with them, even though an efficient update may require moving away from the policy positions the speaker seeks to advance when the speaker’s arguments do not resonate. Agents who learn this way are responsive to the explicit evidence of irrationality and in that sense are committed to the normativity of rationality, but their failure to learn efficiently creates space for the meaningful role of deliberation as argumentation that was a key challenge for a classical rational actor model. Inefficient learning that makes this possible is closely linked to a wide range of prominent behavioral observations by social and cognitive psychologists. Recent experimental work offers strong explicit support for it as well, as it does for one of its key strategic implications: the more we expect individuals to have difficulties with indirect inferences—and so, the greater the informational value that deliberation likely has for them—the more we should expect them to want to make their best arguments to others. This implication has a counterpart: difficulties with indirect learning may have the effect of increasing the level of informedness on the part of the citizens because they provoke greater argumentation from the speakers, who are less concerned about the possible downside of alienating the audience. Taken together, these results suggest that what might, at first glance, appear to be impediments to learning may, in a strategic setting, lead to the seemingly paradoxical result noted earlier: better, rather than worse, informed citizens, and ultimately, better, rather than worse, policy decisions. From the standpoint of democratic theory, this is clearly good news, and it suggests that the strategic implications of how individuals respond to argumentation may be a critical factor in bridging the informational and cognitive demands placed on the citizens by the normative theories of democracy and the more modest expectations that are urged by social and political psychologists.

The same model of individual learning that underlies these results gives rise to ideological polarization, confirming the association of polarization with deliberation. Polarization through deliberation comes about because individuals who are not great at making indirect inferences from unpersuasive arguments learn more (or, put differently, receive more information
their own prior biases, and in learning from those sources, they learn in a way that strengthens and accentuates those prior biases. Starkly, polarization, which emerges from this model as a product of inefficient learning, is a consequence of the same model of learning that generates the positive informational effects discussed above. What is responsible for the differences between these seemingly disparate results is the individual decisions on how (or where) to acquire information. If given a choice, individuals will seek information from sources from which they are likely to learn more, and that will lead to the reinforcement of their biases.

If the explanation suggested by this model captures an empirically important part of the empirical story of polarization, it will not do to tell citizens that they should also listen to the other side—if their time is finite, they would prefer not to do that. In fact, from the standpoint of social welfare, it may be better if they did not listen at all and went with their prior—uneducated, but relatively unbiased—judgments. But this model of polarization also points to some possible ways of neutralizing it. One such way is through increasing the influence of decision-making bodies, such as citizen juries, that do not make information acquisition choices on their own but are presented with and must process in real time a cross-section of arguments. While there are good reasons to be cautious when extrapolating from the results of deliberations within such groups, their structure illustrates principles of deliberative design that may be effective in minimizing the downsides of inefficient learning in deliberation. Another approach is by institutionalizing an explicit right of rebuttal. Once an argument for a given policy choice is already made, the downsides of rebutting it have gone down—those who were going to be turned off by it were likely just drawn to the position defended by the opposing speaker. Arguments on behalf of a given policy proposal, thus, encourage arguments against it, and the opportunity to rebut an opposing argument should be expected to lead to real argumentation rather than mere posturing.²⁹

IV

In a textbook formulation, models of rational decisions have a curious status—they are both normative and positive at once: normative in suggesting what the best choices are and positive in describing the behavior of the relevant individuals making those choices. This duality of rational
decision theory makes it an immediately relevant tool of both normative analysis and behavioral social science. However, in that duality there is also a tension: greater emphasis on behavior has traditionally meant marginalization of the normative, and the focus on the strategic settings, which are central to politics and in which predicted behavior is often socially suboptimal, has brought the analysis of rational behavior away from the immediately normative implications of rational decision theory. It is, perhaps, not surprising, then, that political philosophers have been somewhat apprehensive about the more behavioral developments in the social sciences.

Behavioral political economy is, in one sense, a move farther away from the normative side of that duality. The individual behavior it posits is not optimal and, as such, certainly not normative. But, of course, the normativity of optimal choices is not the only sense, and surely not the most interesting of the senses, in which the political economy analysis is relevant to the philosophical arguments. By far the more important sense is that of informing the normative argument of the potentially unanticipated consequences that come with strategic behavior in different institutional settings. And it is here that BPE may hold for political philosophy a much greater value.

To buttress the normative inquiry, political economy models must be able to shed light on the behavioral practices that are central to the normative arguments. Some of the most important of such practices involving citizens’ deliberative activities present a special challenge to models based on the classic rational actor assumptions but can be effectively analyzed in BPE models that share the core methodological and philosophical values such assumptions represent. Those models and the broader BPE approach have a lot to offer to democratic theorists as attractive tools alongside classic rational actor models of other normatively relevant phenomena.

**Dimitri Landa** is Associate Professor of Politics at the New York University and former National Fellow at the Hoover Institution at Stanford University. He has authored a number of articles in leading political science, economics, and philosophy journals.
NOTES

The arguments discussed here draw on many years of deliberation and joint work with Cathy Hafer and have benefitted from feedback and discussions with many political economists and philosophers over that time.

1. In keeping with the view that the differences between the two fields are, in a critical sense, sociological rather than substantive, I will, throughout, use “political economy” as the preferred term of reference to the field that is also sometimes referred to as “formal theory” or “positive theory” and will use “political philosophy” as the preferred reference to the field also known as “normative theory.”


3. Examples of other recent BPE developments that fall under this description include Jonathan Bendor, Daniel Diermeier, David A. Siegel, and Michael M. Ting, Behavioral Theory of Elections (Princeton: Princeton University Press, 2011); Pietro Ortoleva, and Erik Snowberg, “Overconfidence in Political Economy,” American Economic Review (forthcoming); and Alessandro Lizzeri and Leeat Yariv, “Collective Self-Control,” Working Paper (2014). These works advance arguments that pertain to different aspects of democratic accountability and political sovereignty, and as such bear directly on issues of interest to democratic theory. Given the space limitations, I will not attempt to draw out these connections here, but the substantive range of these works underscores the different issues and concerns animating the BPE scholarship.


7. Landa and Meirowitz, “Game Theory and Deliberative Democracy.”
10. See Landa, “Rational Choices as Social Norms.”