Picking the *Right* Players: Applying the Perceptive Interpretation of Game Theory to Rational-Choice Institutionalism

Tomer Perry[1]
Stanford University

Abstract. *This paper examines the choice of players and strategies in the application of game theoretic models to political institutions. Drawing on Ariel Rubinstein's perceptive interpretation of game theory, I argue that the choice should be guided by the principle of relevancy; models should include only factors which are perceived by the players to be relevant. Extending Rubinstein's concept to the choice of players and strategies, I interpret the literature on democracy as self-enforcing equilibrium and argue that the criterion I provide can help settle an argument between researchers in the field regarding the relevant players. Moreover, the approach provides criteria which define what counts as empirical evidence in support of a proposed model.*

"…which, when Anacharsis understood, he laughed at him for imagining the dishonesty and covetousness of his countrymen could be restrained by written laws, which were like spiders' webs, and would catch, it is true, the weak and poor, but easily be broken by the mighty and rich. To this Solon rejoined that men keep their promises when neither side can get anything by the breaking of them; and he would so fit his laws to the citizens, that all should understand it was more eligible to be just than to break the laws."

(Plutarch's Life of Solon, translated by John Dryden)

This paper examines the political science approach called 'new economics of organizations' or 'rational-choice institutionalism' (Weingast B. R., 2000; Moe, 1984) from a methodological perspective. Drawing on concepts developed in economics, this approach is characterized by the study of political institutions with applications of game theoretic modeling and has been applied widely to the study of political institutions (Bates, Greif, Levi, Rosenthal, & Weingast, 1998; North & Weingast, 1989; Weingast B. R., 1998; Weingast & Marshall, 1988; Acemoglu & Robinson, 2000; North, Wallis, & Weingast, 2009; Przeworski, 2004; Fearon J. , 2006; Weingast B. R., 1997; Milgrom,

[1] I would like to thank Foivos Karachalios, Colin McCubbins and Ruth Kricheli for helpful comments on this paper.

North, & Weingast, 1990; Bates, 2001).[2] These examples share basic features of rational choice institutionalism – understanding political institutions as self-enforcing equilibria, which are sets of corresponding choices of actions of players working in service of their preferences. The breadth of applications, however favorable to the generality of the method, raises a concern – how should the approach be applied in specific cases? How should the theory be empirically tested? What sort of data should count as evidence in support of the theory? This paper will examine these questions in the particular terms of the method – how do we determine the game in each situation? What makes a good choice of players and strategies? Drawing on Ariel Rubinstein's (1991) perceptive interpretation of game theory, I argue that the choice of players and strategies should be guided by the principle of relevancy - "a game theoretic model should include only those factors which are perceived *by the players* to be *relevant*" (919, original emphasis).

Expanding Rubinstein's notion to the choice of players, I discuss treating collective agents as players in game theoretic models. I argue that that a group can be treated as a player in a game theoretic model *if features in the considerations of relevant decision-makers as such and if it can be depicted as acting in light of a consistent set of preferences*. Using this approach, I interpret the literature on democracy as self-enforcing equilibrium and argue that the criterion I provide can help settle an argument between researchers in the field regarding the relevant players –whether they are ethnic- or class-based groups. Moreover, I discuss strategies and how they become relevant to decision-making, and argue that strategies in a game theoretic model should be those *possible plans of actions perceived by the actors as relevant alternatives along with the considerations that support the optimality of such plans*. Particularly, I show that there many instances where 'off the equilibrium path' strategies, which are never played according to the models, are 'played' once or twice before they become a viable alternative in the minds of the relevant players. Finally, I claim that the criteria I provide facilitate the application of game theoretic models to the development of

[2] For a different approach to political institutions which my paper does not consider see (Diermeier & Krehbiel, 2003).

political institutions, which critics have claimed to be a weakness of the rational choice research program in political science.

The next section of the paper presents rational-choice institutionalism briefly and the criticisms that it has been subjected to. The third section presents and extends Rubinstein's account of the interpretation of game theory. The fourth section applies the logic to political institutions and exemplifies it using the literature on self-enforcing democracy; the last section concludes and draws some implications regarding the kind of historical evidence that should can be provided in support of a particular model.

Rational-choice applied to political institutions

Rational-choice is a theory of strategic interaction, where autonomous agents (typically individuals, sometimes collective agents) act in order to promote their interests. It is based on the understanding that the social reality is the outcome of many people's actions; people know this and act accordingly, based on what they think others will do. Ascribing preferences and autonomy is, respectively, the basis of 'rational' and 'choice'. A set of such corresponding choices, made by agents who are aware of the others' decisions and respond choosing the best alternative, is called an equilibrium. A crucially important feature of equilibria is that they are *self-enforced*, that is they are brought upon and reinforced by the strategic actions of players.

In this framework, institutions are solution to collective action problems (North D. C., 1990; Weingast B. R., 2000). The setting of autonomous individual decision-makers is characterized by lack of cooperation. More particularly, individuals face commitment problems which preclude them from enjoying possible benefits of cooperation (Williamson, 1985; North D. C., 1990; Weingast B. R., 2000). There are benefits to cooperation but they are thwarted by incentive problems: each individual is tempted to defect for personal gain. If only they could agree to forgo personal gain, everybody would be better off. However, without institutions agents cannot make *credible*

commitments to cooperate; they all know each of them will have something to gain by defecting. Institutions restructure incentives to make cooperation possible – if you are punished when defecting, for example, your commitment becomes credible. Analyzing institutions as solutions to such collective action problems is the hallmark of rational-choice institutionalism.

This summary, I think, captures the strength of rational choice theory: it offers a simple logic, a single theoretic scheme that can be applied in various different settings. This parsimoniousness is greatly valued by proponents of the approach as well as critics.

It is important to note the critical intellectual environment in which rational-choice institutionalism developed, since it played an important role in shaping the theory. It is through the criticism of competing theories that a theory overcomes its weaknesses; by striving to accommodate such criticisms, a theory transforms into a dominant paradigm that is foundation of a mature science. Rational-choice institutionalism developed in a wave of renewed interest in the study of institutions in political science (Hall & Taylor, 1996). Though prominent in the field, it has been criticized repeatedly and severely and has developed in light of this criticism, in my opinion to the betterment of political science generally. I will mention here two salient criticisms. The first is that rational-choice in political science has focused on developing abstract theories, remaining removed from evidence and empirical tests (Green & Shapiro, 1996). The second, coming mostly from historical institutionalists, is that rational-choice fails to 'take time seriously' and provide an account of institution development; a related point is the focus of rational choice theorists on 'micro' rather than 'macro' analysis (Hall & Taylor, 1996; Skocpol, 2000; Skocpol & Pierson, 2000; Streeck & Thelen, 2005; Thelen, 1999; Pierson, 2000). These criticisms are related – and they concern the core of rational choice, namely game theory. Nevertheless, it is no less important to note that rational theorists in political science, and most of the works of this strand which are cited in this paper, have responded to these criticisms, attempting to integrate insights of historical institutionalism as well as applying their theory to various different empirical explorations (Bates,

Greif, Levi, Rosenthal, & Weingast, 1998; Greif & Laitin, 2004; Weingast B. R., 2000; Weingast & Katznelson, 2005). These efforts, however valuable to the development of the field, are related to the general question dealt with by researchers of other fields – how should game theory be applied to reality? This question, I believe, requires a general answer. And to answer it, I draw on a prominent game theorist who tackled it.

Following Rubinstein's stance – on the interpretation of Game Theory

Rubinstein (1991) focuses on the relations between theory and reality in the field of game theory. He explores the notions 'game form' and 'strategy' generally, but the games he has in mind are mostly played by single players and under fictitious general circumstances that are meant to represent many different social phenomena, such as the prisoners' dilemma or battle of the sexes. Adopting his scheme, therefore, would require adjusting it to the particulars of political institutions. Namely, it will require addressing the question of collective actors – an issue that is too broad for the purposes of this paper but one I will try to take on briefly. Game theory is a specific type of modeling; it concerns the strategic decisions players make in light of decisions made or about to be made by other players. While methodological discussions about modeling are relevant, I narrow my claim and focus on game theoretic modeling. I believe that game theory is especially suitable for analysis of political interaction, as strategic calculations are characteristic of political actors who pursue what they take to be ultimate goals and are usually aware that these goals are not shared by all relevant[3] parties. These two features – the significance of one's own goal and the awareness of other players' attempts to achieve a different goal make game theory especially suitable for the study of politics in general and political institutions in particular. These features, though encouraging of the application of game theory to political settings, also support, I hold, adopting an interpretation of game theory that is along the lines of that proposed by Rubinstein.

[3] By 'relevant' I mean here all parties which are conceived by the acting player to have an influence over the relevant decision. What that means is exactly the topic of Rubinstein's article and following him, my paper.

Rubinstein purports to define two basic terms of game theory – 'game form' and 'strategy'. His two suggestions fit into the same general picture which says that a good game theoretic model "is not meant to be isomorphic with respect to 'reality' but rather with respect to *our perception* of regular phenomena in reality" (910, emphasis added). When we are modeling perception, we focus on how players see the interaction they are pursuing – this is a criterion to select relevant strategies among the theoretically infinite possible strategies. I will extend this logic to the choice of players in the design of a game model, as will be discussed below. Rubinstein does not extend his logic to the choice of players but he does supplement the traditional definition of strategy as a 'plan of action' when he defines it as "a player's plan, as well as those considerations which support the optimality of his plan" (910). This definition stems from and is consistent with some standard game theoretic interpretation, such as that of extensive form games which represent sequential interactions. In such games a 'complete strategy' includes decisions made by a player in every node of the game, including such nodes of the game that would not be reached due to that player's own action. For example, a player who intends to accept any offer in the first phase of a continuous bargaining game is still 'required' to specify his actions in subsequent rounds. Such decisions are important – even though they are 'off the equilibrium path', they are an essential part of the equilibrium, since other players are playing according to their beliefs about such decisions. Consequently, Rubinstein explicates his definition for extensive form games, "a strategy encompasses not only the player's plan but also his opponents' beliefs in the event that he does not follow that plan" (911). Thus, a player's subsequent moves are included as part of the *opponent's strategy* in an equilibrium. And so, the same action figures in the equilibrium (and indeed they are required for the determination of equilibrium) but they figure in it through the strategy of a different player.

This may seem like a linguistic quibble, but I believe it is not. Rubinstein builds on a notion of relevance, which serves as an instruction for modeling, "a game theoretic model should include only those factors which are perceived *by the players* to be *relevant*" (919, original emphasis).

Rubinstein uses his notion of relevance to offer a solution to a case discussed by Eric van Damme and other economists,[4] where the solution[5] of a game changes when one player has an arbitrary ability to discard one dollar prior to the game and that fact is known to the other player. Rubinstein argues that such disposal of a dollar is irrelevant to the strategic considerations players are making in the course of the game and therefore leads to an irrelevant result. This emphasizes the divergence between the model and "a rigid description of the physical world" (919) – it is true that a player can discard of a dollar, and that he can also make sure the other player knows it. But, although there are many things people can do to one another when they interact in cases which we would like to model in game-theoretic terms, how should we know which of these courses of actions should be included in the model? Rubinstein's answer is – those which the players themselves see as relevant to their decision-making should be included.

The point generalizes. As I mentioned, I take this point to apply to the determination of players in the framework of the game form. In many cases, it is clear who are the players involved – such as the case when there is a set number of parliament members (Weingast & Marshall, 1988). However, in many other interactions, it is physically possible for many other people to interfere. When a buyer and a seller negotiate a deal, there could be many other people who are willing to make counter offers that can obviously change the dynamics dramatically. If the bargaining sides do not see these outsiders as relevant to the deal, they should not be included in the strategic analysis of the situation, modeled in game theoretic terms. Conversely, the applicability of the game depends on players making such considerations when they are making such decisions. For example, Fearon and Laitin (1996) present a social matching game where individuals are paired randomly in

[4] See Rubinstein (1991) for the references.

[5] This is a problematic statement, since van Damme is using successive elimination of weakly dominated strategies, which is not a standard solution concept that is widely accepted. However, I use this example since Rubinstein uses it – I'm assuming that it's possible that such an arbitrary modification of a game can change the solution of the game even if we limit ourselves to more accepted solution concepts.

successive turns to play a prisoners' dilemma, meeting either a coethnic or a person from a different ethnic group.[6] To explain the situation, Fearon and Laitin say:

As an example of what this set-up is intended to represent, one may imagine a number of traders who go to a market each day, meet other traders, arrange or fail to arrange deals, and return the next day to interact with new partners (probably). Or one may imagine that each day people wander out into the world and have chance social encounters that may be good or bad, depending on how both parties act. "Defection" can be interpreted as an attempt to cheat or rob the other player, or in some contexts can be taken as avoiding or forgoing the interaction (719-20).

The game seems to capture well the market set-up – where all traders come, meet and mix, make deals and have short-term incentives to cheat one another. At the same time, the idea that the prisoners' dilemma represents generic 'chance social encounters', capturing the strategic essence of people's experience with the categorization of 'good' and 'bad' which determine whether or not they will want to repeat this interaction seems too crude: it abstracts away from important strategic considerations which guide people's choices. Fearon and Laitin acknowledge that too, and say "The game G modeled interactions in which all individuals have an incentive opportunistically to exploit their partners. For many interactions, however, such as routine encounters in the street, it is more realistic to suppose that people vary in their disposition to behave opportunistically (or aggressively)" (726). However, this concession raises a question – if 'routine encounters in the street' are not well-captured by the prisoners' dilemma, why should we accept it characterizing 'chance social encounters'? When does the model apply? My answer is that it applies in those settings where people make such considerations. Therefore, the marketplace is indeed a quintessential arena to examine the predictions of the model. Furthermore, since Fearon and Laitin

[6] Fearon & Laitin's model only includes two ethnic groups – which raises the same question, since their examples include religious affiliation as a type of 'ethnic' group. How do we know how many groups we have in the society? for a game-theoretic model, I would say, the answer depends on the number of ethnic groups that figure in people's decision-making of the sort we are analyzing.

aim to discuss the model in terms of two ethnic groups – I would argue that it holds for decisions made by people with regards to issues where ethnicity is held to matter. So, for example, if ethnicity is not generally held to matter with regards to the ability of a dentist, such considerations will not feature when one is choosing a dentist and the predictions of the multi-ethnic game would be wrong, *even if* other features of interactions are similar to those of a prisoners' dilemma and the dentist is of a different ethnicity from the client. These qualifications, I believe, do not weaken the model – in fact, they strengthen it by sharpening its focus and enriching the empirical demands it makes on the scenario, as well as allowing for more precise empirical testing.[7] The next section will discuss the implications for empirical application of game theoretic models. My point here is not about ethnic groups and the generality of prisoners' dilemma in social settings; indeed, the claims I made about these here tentative. The point is about the applicability of game-theoretic models.

An additional point is that the choice of players is part of the strategic understanding of the situation. This means that over-inclusion is also a mistake from a methodological point of view. Including in the model players who may influence the game but are not perceived as such by other players would yield incorrect predictions. Imagine a game played by two super powers, the US and Russia. They consider each other's moves regarding the Russian (or American) invasion to Afghanistan. However, it is possible for Pakistan to interfere in the situation by sending a platoon on a suicide mission. If such a platoon is sent, that would change the military balance significantly and if either the Russians or Americans considered their actions in light of such intentions, it would impact their decision-making. Nonetheless, both reject this possibility as highly unlikely, assign a zero possibility to any such occurrence and devise their strategies against one another as if there is no Pakistani player. Including Pakistan as a player in the game analysis would be a mistake – it would lead to the wrong predictions regarding the moves both the US and Russia would make or,

[7] Part of the problem of empirically testing a theory that purports to suggest a general rule is that anomalies can be explained ad-hoc as exceptions or results of some unknown environmental variable which does not defy the general rule but impedes its implementation in the specific situation (see the discussion in Green & Shapiro 1996, 180-183).

alternatively, it would explain the mechanisms according to which they reached their decisions wrongly (assigning to them the wrong preferences, for example). Furthermore, even if we are analyzing a game which concern events that have already occurred and thus know that Pakistan has indeed sent their suicide squad, this does not mean that either the US or Russia made the right predictions and were considering such a move by Pakistan as a relevant consideration when devising their strategies.[8] If we include the Pakistani player, we fail to capture the strategic interaction. However, as I will discuss below – players are not stupid and they may learn from their mistakes. In case Pakistan does interfere, the US and Russia might later include a third-party player intervention *even when such a player is not even considering intervention.* The logic, I hope, is clear and leads us to the kind of empirical evidence that is required by the application of game theory to political institutions which will be discussed in the next section.

But before I go on, I would like to clarify the point. The relevancy criteria and the preceding discussion seem to push me to accept that game theoretic models should include players that do not exist.[9] I may think that there is a person behind the door and act accordingly, or believe that the shadow I see is that of a person though it is only a tree shaking in the wind. Should we include the tree as a player in the game because it features in my considerations to dock and cover? In part, I would bite the bullet – yes, I think we should model the tree as part of the game. But that does not mean we need to include the tree as a player, for the tree does not make any decisions, though I may think it does and act in light of what seem to me to be a decision made given alternatives (I may think 'if he comes any closer than the lamp-post despite my warnings, I shoot!'). We would have to feature the tree in the game as we do with other chance events (usually consider them as 'nature' players which make stochastic decisions) and model the misinformation that I have regarding the nature of the player 'nature'. There would be a different though – for although the

[8] I hope to discuss separately the problem of over-estimating the probability of an event post-mortem and its significance for social sciences.

[9] I discuss this issue generally here for it will feature again later in the discussion of groups as collective agents.

tree would be strictly playing a mixed strategy that is best captured by a stochastic process, I would still be considering it as a strategic player. That is, I would assume it is playing according to strategic considerations – avoiding strictly dominated strategies and so forth. But that is fine – with Rubinstein's definition, I take it that these beliefs about the tree's behavior would be considered part of my strategy. Game theorists have developed ways to deal with such cases where one or more players are unsure about the other player/s and/or their strategies.[10] While it is a peculiar case, it does not seem be beyond the capability of modeling – it only requires that we somehow connect sets of results the a stochastic process (nature's decisions) with the 'decisions' of the player I mistakenly believe exist.[11] This case is more likely to occur, I believe, with regards to collective agents, which tend to feature in game-theoretic models of political institutions.

Applying the logic to institutions

In this section I will discuss the choice of players and strategies in the context of political institutions, on the basis of the analysis provided thus far, using examples from the literature.

Players

Who are the relevant players to an interaction? Given the autonomy we attribute to players, it is somewhat natural to think of the players as individuals and indeed some applications focus on individuals' actions (Weingast & Marshall, 1988). However, in some cases, models include collective unified actors as decision-makers and indeed, where the focus is on 'macro' analysis such attribution is common and useful in simplifying the case. The issue concerning the possibility and

[10] There is also work done regarding the cases where different players perceive differently the games they are playing. An interesting and unique example is Yossi Feinberg's work on the concept of 'unawareness' (Feinberg, 2004).

[11] Consider the case where I see a snake in a cage and wonder whether it's a real snake or a mechanic one. If the snake is mechanic I would want to reach into the cage (say there is a precious gem in it) but if it is real I would prefer to wait until the snake is examined or removed by an expert (but if it isn't real, I wouldn't want to wait for the expert, nor to have to share the gem with them!). Assume that a real snake is a simple player which reacts in predictable ways (if you throw a rock near it, it recoils and if you throw a rock at it, it attacks). We can model the situation by assigning a probability to the possibility that the snake is fake and solve the game for the decision-maker (me). We can then also allow that probability to change according to the interaction, allowing me to update my beliefs (perhaps using Bayes law) regarding the fakeness of the snake. This kind of model, I think, could apply the tree situation and could generalize to a more sophisticated player than the snake.

plausibility of collective agents is a complicated one which many have worked on (Olson, 1971; List & Pettit, 2011; Searle, 1995). List & Pettit (2011) argue that groups count as collective actors if they can act in accordance to a consistent set of attitudes. Without delving into the intricacies of their argument, I take their definition to be suitable for the discussion of game theory. A collective agent is a player in the relevant sense if it can be seen to make decisions according to a consistent set of preferences. However, this does not help us much,[12] for it merely is the definition of a player in game-theoretic terms. Indeed, this is what we are assuming about groups when we treat them as agents. The question remains, when can we assume that groups can make such decisions?

I eschew the deeper philosophical question with my focus on game theory – the question here is not whether or not groups are indeed collective agents in any metaphysical sense but whether or not it makes sense to model them as such in game-theoretic settings. I argue that it does, though it does emphasize the difficulty presented by the choice of players in a specific game. Given my previous discussion of the interpretation of game theory, my answer is that *we should treat a group as a collective agent in game-theoretic models if such a group features in considerations of relevant decision-makers and if it can be seen as acting in light a consistent set of preferences*. These are the conditions to the applications of the model, as I will argue, and evidence in their support should be seen as evidence in support of the adequacy of the model to a particular situation or alternatively, when they hold, as the proper grounds for testing the model.

But what if all players are groups? And what if a group is conceived as an agent by other players but not by the members of the group itself? For example, the parliament may extend the franchise since its members are afraid of 'the poor' starting a revolution, while the poor people may not see themselves as a collective agent capable of such a feat. Or conversely, some people among the poor may think that they, the poor, are capable of leading a revolution while the rich elite may dismiss this idea as nonsense and not consider it as part of their strategic calculations. These cases,

[12] List & Pettit say much more than that about the conditions of group agency and their argument can help us, as will be discussed below.

I believe, could be resolved using the same form of modeling that was referred to previously in the example of the frightening tree shadow. The interesting cases, where game theory would have something to teach us, would be those where there are multiple decision-makers who are aware of one another and respond to one another. This is true simply because the machinery of game theory is geared towards attending to these kinds of situations. Nonetheless, these difficulties show that those different situations will require tuning the tools of game theory to the specific situation. If the parliament takes peasant riots as signals from a collective player they perceive as 'the poor', while the peasants riot in response to wheat prices – the game should be modeled differently than if the peasants are expressing their anger at the ruling class. The fact that the peasants do not see themselves as part of a collective agent is not a decisive case against them being one – it may be the case that they will be a collective agent in List & Pettit's terms by following blindly a small group of political or religious leaders. While they do not perceive themselves as a collective agent, their collective actions can be described in terms of a consistent set of preferences. List & Pettit (2011) discuss shortly the possibility of group agents without joint intention (24-5) and suggest that this can happen in two ways: either by a process of cultural or natural evolution or when "one or several organizational designers co-opting others into a structure underpinning group agency, without making them aware of their agency at the group level and without seeking their intentional acquiescence in the arrangement" (24). The former way, however, is unlikely to bring into existence new group agents (though it would probably shape them once they exist). The latter way, though not discussed at length by List & Pettit, seems to be highly relevant to historical examples that feature in the literature, and so I now turn to these.

Consider the models for self-enforcing democracy (Weingast B. R., 1997; Przeworski, 2004; Fearon J. , 2006; Acemoglu & Robinson, 2000). Weingast (1997) models the citizenry as two groups of citizens, dubbed A and B. But who are these groups? They take on different identities throughout the article and, in fact, it seems that they can be ethnic groups in a divided society as well as

coalition groups characterized by different economic interests, such as the Tories and the Whigs in 17th century England. Fearon (2006) says his model can be an *n*-player version of Weingast's, having a large number of citizens instead of two groups. However, Fearon allows for his citizens to stand for groups and also admits that this makes a difference – with smaller *n* there are some equilibria which seem much more plausible than others.[13] If it matters, then, how are we to know? Fearon hints at a condition for treating a group as a unified, but he only mentions in passing while saying "One possible response is to argue that it makes more sense to take the 'citizens' in the model to be *groups that have already overcome internal collective action problems*" (14, emphasis added). Thus, a group of individuals can only be treated as unified agent only if they solved a *different* collective action problem, the solution of which is a precondition for assuming they are a player in larger scale.

To see why this matters, consider a different strand of accounts of the idea of democracy as a self-enforcing equilibrium which bases its analysis on players defined in terms of class, by their economic interests or standings (Acemoglu & Robinson, 2000; Przeworski, 2004). Przeworski (2004) rejects the idea that ethnic groups are unified actors and focuses on games played with two players which he dubbed 'the poor' and 'the rich'. In principle, it seems that such games could be captured by Weingast's general scheme, which can accommodate any two groups. Tying the group-players to economic interests allows Przeworski to make predictions regarding the sustainability of democracy under different conditions of economic growth and income inequality.[14] But why are we to treat economic classes as unified players? Apart from the fact that results fit empirical data, Przeworksi doesn't supply an answer. However, he also mentions in passing an idea that may hint at what should probably be a condition for treating a group as a collective agent when he says, in

[13] The idea being that it is much more likely that an organized group can protest in a way that will become common knowledge to a small group of other groups than a single citizen make a protest that will become common knowledge to all, or most, other citizens. Weingast (1997) relies on the same idea when he says "most societies are unlikely to resolve the coordination problem in a wholly decentralized manner" (251).

[14] Of course, Przeworski's model is largely influenced by the results of his empirical research and his model aimed at explaining such results. If we treat Przeworki's work as an attempt to 'fit the data' with a model, we may not see the results as 'predictions' at all. I assume that we do want to see it as predictions and that drives my analysis.

the opening passage of the article: "Hence, to survive, democracy must be an equilibrium *at least for those political forces which can overthrow it*" (1, emphasis added).

Both these remarks, I hold, are covered by my criteria for identifying players in a game-theoretic model as group agents, in the sense explained, that are relevant to each other's decision-making. Moreover, my account systematizes them in a one unified principle. Fearon's remark alludes to the fact that groups have collective action problems and would therefore not be able to act in light of a consistent set of preferences. Przeworski's comment on the ability of a 'political force' to overthrow the reigning political force can either be interpreted in a similar fashion as a group's inability to act in light of a consistent set of preferences, due to a collective action problem or for any other reason. However, his remark can be interpreted in terms of the groups that are perceived to be relevant as 'political forces'. Despite their disagreement about the relevant groups to the sustainability of democracy, Fearon and Przeworski both appeal to a condition of group agency though neither of them provides one. I believe my condition generalizes their concerns and allows for adjudication between the competing theories – it allows examining whether relevant actors in a particular case are ethnic or class based collective agents. The predictions of these models are similar and we cannot choose between them; the criteria I provide helps supporting one model against the other, suggesting which of the mechanisms captures best the interaction depicted by the historic evidence.

Strategies

The logic I will argue for determining strategies is similar to the one presented in the previous section concerning the determination of players and consistent with Rubinstein's definition. Strategies in a game theoretic model should be those *possible plans of actions perceived by the actors as relevant alternatives and the considerations that support the optimality of such plans*. The determination of strategies, although theoretically separate from the determination of players, may commonly be accompanied with the determination of players: showing that a group can act in

light of a consistent set of preferences would regularly include suggesting some forms of actions which can then be the basis of sketching the set of alternative strategies. However, determining that a particular group counts as a collective actor does not settle the question of possible alternatives. Considering the cases discussed so far, if it can be shown that the class of the relatively poor can be seen as a collective agent, that is they can be seen as acting in pursuit of a consistent set of preferences, it does not answer the question whether or not they are capable of, for example, overthrowing the ruling class. Even if it can be plausibly shown that they are indeed capable of such a feat it does not necessarily mean that they, or any other relevant players, see it as a relevant consideration in their decision-making. The poor may not be aware of their capacity or may not believe in it (Feinberg, 2004). Likewise, others may dismiss them as incapable.

Acemoglu & Robinson (2000), for example, argue that in some western countries in the 19th century the political elite extended the franchise to prevent a revolution. As evidence that such was the case in Britain, Acemoglu & Robinson cite the British Prime Minister Earl Grey saying in the parliament in 1831 that "the principal of my reform is to prevent the necessity of revolution" (quoted in Acemoglu & Robinson, 1182). While this quotation certainly supports their argument, it fails to explain why the franchise was extended in 1832 and not earlier. What made it the case that the threat of revolution became credible around that time? Surely it is not the mere capacity of the masses to overthrow the government – for they had the same capability, technically, fifty or one hundred years earlier. It is possible that the answer has something to do with their ability to act as an organized group, but I do not wish to present a historical argument here. I do want to emphasize – what matters to the type of model that Acemoglu & Robinson present is that the political elite *perceived* the poor as a collective agent capable of a revolution. I argue that what is important about this evidence is that it shows that the possibility of revolution featured in strategic consideration of the parliament when they were debating the extension of franchise.

There is another interesting complication here. The model presented by Acemoglu & Robinson predicts that the political elite will extend the franchise in order to avoid a revolution, meaning in order to obviate the need for a revolution on the part of the poor. They present a game with two players – the rich and the poor.[15] The poor player can overthrow the government, as was discussed earlier, and it is the threat of such revolution that drives the result, which is that the political elite (in this game, the rich player) extends the franchise. However, according to this game there should be no revolution – it is an 'off the equilibrium path' strategy which is never actually played. Such strategies are significant for they determine the equilibrium – if we assume that the poor cannot overthrow the government or, alternatively, that the rich do not perceive them to be capable of it, the equilibrium would change and there would be no extension of the franchise. Nonetheless, 'off the equilibrium path' strategies present possible plans of actions that do not materialize, they are at best credible threats/promises. If this is the case, how can it be that Acemoglu & Robinson cite "unprecedented political unrest" (1182-3) and a series of riots as evidence in support of a model that predicts that such unrest and riots would be avoided by the extension of the franchise? The answer is that the occurrence of such riots supports not only that the poor were *capable* of a revolution but they also made it the case that the rich realized that a revolution is a viable alternative. These events changed the way the rich *perceived* the game, the way they understood the player they stand against and the alternative strategies it may employ in service of its preferences. This, I argue, is what makes the predictions of the model accurate in the beginning of the 19th century and not at the end of the 18th century – the rich started seeing the possibility of a revolution as an alternative strategy of the poor, and that has started to feature in their own strategic considerations.

[15] This is not entirely accurate – they have an infinite horizon of agent, a proportion of which are poor (the rest being rich). However, all poor agents are identical and so are all rich agents and they act as one agent and the fact they are many is only important since the proportion of the respective populations is a parameter of the model (see p. 1170).

I believe this is not a coincidence. 'Off the equilibrium path' strategies, which in theory should not occur, are often 'played' at least once or twice as to establish themselves in the minds of relevant players as viable alternatives. Indeed, this is one way in which threats and promises become credible, which is a core concept of rational choice institutionalism. This point is made by North and Weingast (1989), in their article concerning the institutional change following the Glorious Revolution, when they say: "part of the glue that held the institutional changes together was the successful dethroning of Charles I and, later, James II. This established a credible threat to the Crown regarding future irresponsible behavior." (816) Thus, the equilibrium following the glorious revolution rested on the existence of a threat, which means in game theoretic terms that it was common knowledge to the Crown as well as the parliament that the latter can dethrone the king, and that threat was able to deter the king from reneging on his loans and violating the property rights and personal liberties of those classes represented in the parliament. Dethroning the king is an 'off the equilibrium path' result – it does not happen in the period following the Glorious Revolution for it is obviated by the Crown's checked behaviour. However, James II, the king dethroned in the Revolution, and Charles I, the king who was beheaded in the civil war – may not have considered dethroning to be a possible plan of action on the part of their opponents. When James II turned against the Tories who supported him, he might have assumed that they will not turn against him since they "maintained a notion of passive obedience to the sovereign, believing in acquiescence in the face of undesirable acts" (Weingast B. R., 1997, p. 252). We see that an 'off the equilibrium path' strategy can come into existence through its previous application, which provides the most promising evidence that the decision makers see it as a possible alternative. This description fits the principle of relevancy I suggested in this paper.

Conclusion

This last point regarding 'off the path equilibrium' strategies, I believe, is an important insight regarding the application of game theory to historical narratives and it follows from its

attempt to rise to the challenge of its critics. If rational choice theorists wish to 'take time seriously' they should give an account of how players and strategies are come to be and develop over time. The criteria I provide for the determination of players and strategies form the theoretical framework that allows rational choice theorists to provide such an account. Moreover, I believe, these criteria support the application of game theoretic models to empirical testing as well as help those models provide an account of institutional change and transition from states of equilibria. The coming into existence of strategies and players is a mystery to the naïve version of game theory – players and strategies are assumed by the model. But we don't know when did a player or strategy come to be and so whether or not we can assume that a player or strategy exists; the existence of such a player or a strategy has an impact on the predicted equilibrium result even if it never plays or is never played. The principle of relevancy and the considerations presented in this paper guide us in the search of evidence – it tells us what historical events support the application of our model. If game theoretic models are to be applied to empirical settings, we need a way to determine which of the many possible game theoretic models that can be applied to a particular situation and have different predictions, should be chosen. Rational choice literature in political science has developed enough for this problem to be salient, as in the case of the different models of self-enforcing democracy. Multiple models with different sets of actors seem to produce the same predictions; how can we adjudicate between them? How can we know which of them capture the essence of the interaction? In the existing literature it is rare to find an argument for the choice of players or strategies – they are treated as assumptions that need not be validated. If I am right, the implications are that without such validation, it is impossible to empirically test the model against others with different sets of layers.

These implications can be seen in two different ways. First, they can be seen as conditions to the applications of our models. We devise a model – suggest players and strategies and calculate an equilibrium which serves as the prediction of our models; that is, it tells us what to expect in

settings where our model holds. Then the conditions I provide tell us what are the settings in which we should expect the predictions to hold – it allows us to test the model against various different cases where such settings occurred, allowing us to falsify our model.[16] Second, they can be seen as the conditions validating the mechanism of the model. In cases where there is evidence that the equilibrium result holds, my criteria provide support that the model represents the mechanism that supports the equilibrium. This allows us to judge the model against other models with similar equilibrium predictions but different assumptions. Both of these aspects, I believe, are beneficial to empirical applications of game theoretic models in political institutions.

[16] By this I do not meant to adopt a naïve conception of Popperian falsification and only to argue that empirical testing can serve to support a theory against competitors and see footnote 5 above and the sources it refers to.

Works Cited

Acemoglu, D., & Robinson, J. A. (2000). Why Did the West Extend the Franchise? Democracy, Inequality, and Growth in Historical Perspective. *Quarterly Journal of Economics , 115* (4), 1167-1199.

Bates, R. H. (2001). *Prosperity and Violence: The Political Economy of Development.* New York: W. W. Norton & Co.

Bates, R. H., Greif, A., Levi, M., Rosenthal, J.-L., & Weingast, B. R. (1998). *Analytic Narratives.* Princeton: Princeton University Press.

Diermeier, D., & Krehbiel, K. (2003). Institutionalism as a Methodology. *Journal of Theoretical Politics , 15* (2), 123-144.

Fearon, J. D., & Laitin, D. D. (1996). Explaining Interethnic Cooperation. *The American Political Science Review , 90* (4), 715-735.

Fearon, J. (2006). *Self-Enforcing Democracy.* Berkeley: Institute of Governmental Studies, UC Berkeley.

Feinberg, Y. (2004). Subjective Reasoning – Games with Unawareness. *Stanford Graduate School of Business, Discussion Paper #1875* .

Green, D. P., & Shapiro, I. (1996). *Pathologies of rational choice theory: a critique of applications in political science.* New York: Yale University Press.

Greif, A., & Laitin, D. D. (2004). A Theory of Endogenous Institutional Change. *American Political Science Review* , 633-652.

Hall, P., & Taylor, R. (1996). Political Science and the Three New Institutionalisms. *Political Studies , 44*, 936-57.

List, C., & Pettit, P. (2011). *Group Agency: The Possibility, Design and Status.* Oxford University Press.

Milgrom, P. r., North, D. C., & Weingast, B. R. (1990). THE ROLE OF INSTITUTIONS IN THE REVIVAL OF TRADE: THE LAW MERCHANT, PRIVATE JUDGES, AND THE CHAMPAGNE FAIRS. *Economics & Politics , 2* (1), 1-23.

Moe, T. M. (1984). The New Economics of Organization. *American Poiltical Science Review , 28* (4), 739-777.

North, D. C. (1990). *Institutions, Institutional Change and Economic Performance.* New York: Cambridge University Press.

North, D. C., & Weingast, B. R. (1989). Constitutions and Commitment: The Evolution of Institutional Governing Public Choice in Seventeenth-Century England. *The Journal of Economic History , 49* (4), 803-832.

North, D. C., Wallis, J. J., & Weingast, B. R. (2009). *Violence and social orders: a conceptual framework for interpreting recorded human history.* New York: Cambridge University Press.

Olson, M. (1971). *The Logic of Collective Action: Public Goods and the Theory of Groups.* Cambridge, MA: Harvard University Press.

Pierson, P. (2000). Increasing Returns, Path Dependence, and the Study of Politics. *American Political Science Review , 94* (2), 251-267.

Przeworski, A. (2004). Self-Enforcing Democracy. In B. R. Weingast, & D. Wittman (Eds.), *Oxford Handbook ofPolitical Econom.* Oxford: Oxford University Press.

Rubinstein, A. (1991). Comments on the Interpretation of Game Theory. *Econometrica , 59* (4), 909-924.

Searle, J. R. (1995). *The Construction of Social Reality.* New York: Free Press.

Skocpol, T. (2000). Commentary: Theory Tackles History. *Social Science History , 24* (4), 669-676.

Skocpol, T., & Pierson, P. (2000). Historical Institutionalism in Contemporary Political Science. In I. Katznelson, & H. V. Milner (Eds.), *Political Science: The State of the Discipline* (pp. 693-721). New York: W.W. Norton & Co.

Streeck, W., & Thelen, K. (2005). Introduction: Institutional Change in Advanced Political Economies. In W. Streeck, & K. Thelen (Eds.), *Beyond Continuity: Institutional Change in Advanced Political Economies* (pp. 1-39). Oxford: Oxford University Press.

Thelen, K. (1999). Historical Institutionalism in Compartive Politics. *Annual Review of Political Science , 2*, 369-404.

Weingast, B. R. (1998). Political Stability and Civil War: Institutions, Commitment and American Democracy. In R. H. Bates, A. Greif, M. Levi, J.-L. Rosenthal, & B. R. Weingast (Eds.), *Analytic Narratives* (pp. 148-193). Princeton: Princeton University Press.

Weingast, B. R. (2000). Rational-Choice institutionalism. In I. Katznelson, & H. V. Milner (Eds.), *Political Science: The State of the Discipline* (pp. 660-692). New York: W.W. Norton & co.

Weingast, B. R. (1997). The Political Foundations of Democracy and the Rule of Law. *American Political Science Review , 91* (2), 245-263.

Weingast, B. R., & Katznelson, I. (2005). Intersections Between Historical and Rational Choice Institutionalism. In B. R. Weingast, & I. Katznelson (Eds.), *Preferences and Situations: Points*

of Contact between Historical and Rational Choice Institutionalisms (pp. 1-24). New York: Russell Sage Foundation.

Weingast, B. R., & Marshall, W. J. (1988). The Industrial Organization of Congress. *Journal of Political Economy* , 132-163.

Williamson, O. E. (1985). *The Economic Institutions of Capitalism.* New York: The Free Press.