Rational Choice Theory as Formalized Common Sense

Luis Fernando Medina

September 1, 2004

1 Introduction

Few methodological issues are more stubborn than the status of rational choice theory (henceforth, RCT): the debates surrounding it are approximately five decades old with no end in sight. Fifty years are a long time not only in the life of any human being but also in the progress of a discipline so it is remarkable that the discussion is not closer to a conclusion now than what it was at its inception. Perhaps it is time for those of us who believe in RCT to accept that, for all our efforts, we may never be able to convince our critics and that RCT may never attain the predominance in social sciences we would have hoped for. The more time passes, the more likely it is that RCT will be swept aside by some newer paradigm without ever having attained the status of an orthodoxy. In the scheme of scientific progress, RCT may some day be perceived as a stillborn doctrine.

Whatever the future holds in store for it, I believe that RCT deserves to be defended although for reasons that other RCT scholars may disagree with. In this paper I will concede many points raised by RCT's critics but I will also claim that, with some reinterpretations, both methodological and substantive, RCT can, in cooperation with other paradigms, clarify several important phenomena and, in the process, deflate some pseudo-problems.

Among the many criticisms levelled agains RCT, I want to take issue with a few. RCT is flawed, critics argue, because it:

1. Assumes that individuals have fixed preferences, in spite of large evidence to the contrary.
2. Assumes that individuals are solely utility maximizers, neglecting the possibility of other goals (e.g. ethical, altruistic, etc.).
3. Assumes that individuals can make decisions independent of their context (e.g. culture, history, social stations, etc.).

In defense of RCT, I will argue that these are not outlandish and hopeless assumptions but that, instead, they are an adequate reply to some of the challenges that social sciences

pose. I will couch this argument first as a general statement about social sciences, and then I will show an example that illustrates how this set of assumptions, when coupled with new game-theoretic techniques, can in fact illuminate some important social phenomena.

2 Beyond the "Friedman Defense."

In an amazingly prescient preemptive move, Friedman (1955) provided a defense of RCT that, in one way or another, still commands the respect of many of its practitioners. Adopting a purely popperian stance on the role of sciences, Friedman argued that in accepting or dismissing a theory, we should only look at its fit with the data, not at the realism (or lack thereof) of its assumptions. Bluntly speaking, he urged us to apply to theories the principle we often apply to sausages: as far as the final product is good, better not to ask about the ingredients.

This is the famous "as if" defense of RCT: however implausible an assumption may seem, it is acceptable as long as humans behave "as if" it were true. RCT practitioners that take Friedman's defense at heart cannot be bothered by lengthy philosophical discussions as the ensuing one. In their view, questions about human agency, ethical behavior and things of the like are irrelevant for the enterprise of RCT because they do not falsify its predictions.

In retrospect, this response seems to have the glibness typical of youth. Back in 1955, RCT was still a young program full of promise that could well afford this brash attitude of asking to be judged only by its results. But time has not been entirely kind to RCT and the results have not all been as conclusive as its early proponents hoped. Interestingly enough, the stumbling blocks in the road have not persuaded such proponents to abandon the theory whereas, by popperian standards, chances are that the theory should have been rejected long ago.

There is a deep problem with this Popper-Friedman view of theories as sources of predictions: theories have to be built before they can be tested. The narrow criterion of falsifiability does not tell us how to develop our theories and, even more importantly, how to modify them when they face problems. Scientific progress results from revisions of previously existing theories, not from opening a casino where everyone can offer all kinds of hypotheses, in the hope of seeing them confirmed by evidence. New and better theories are possible only after a judicious evaluation of the extant paradigms. Even if RCT's unrealistic assumptions can form part of a good theory, they must be subjected to serious discussion. "As if" arguments certainly have a place in RCT and, I will argue, they are often useful and acceptable, but not solely by virtue of their alleged predictive power. If we are serious about revising and improving our theories, (and, let's face it, RCT could use improvements), we need criteria on how will we recognize a useful and acceptable assumption *before* we test it.

3 Social Sciences and Common Sense

I want meta-theoretical criteria to assess the assumptions of RCT and, to that end, here I will propose and defend a very controversial doctrine which, to my knowledge, is latent in some of the literature, especially in RCT, but rarely expressed in so many words: social sciences must be a language for our common sense.[1]

Before defending this thesis, let me clarify some of its contents. When I say that social sciences must be a language for our common sense, I am not saying that they must mirror exactly our experiences in everyday life. Painting and poetry are languages for our emotions and perceptions, but this does not mean that pictures and poems must reflect all and only our mental states. Likewise, if we accept that social sciences must be a language for our common sense (something I have yet to defend), this does not mean that social sciences must only use categories immediately accessible through common sense. Social sciences are replete of notions we do not encounter in our everyday life (e.g. modes of production, utility functions, subsystems of social integration, etc.). None of this is illicit within the view I am proposing. It may well be that those notions are necessary to articulate our common sense.

What *is* ruled out by this view is that social sciences may overturn fundamental aspects of our common sense. This runs counter to a venerable view according to which the role of social sciences, just as that of natural sciences, is precisely to offer us truths that stand taller than our commonsensical perceptions. The heliocentric theory challenges our everyday experiences but we defer to it. Our senses are wrong in assessing the right movement of the Earth and the Sun, and we gladly adopt a better theory that contradicts them. Why then, should we not apply the same standards for the social sciences?

To answer this question, I need to discuss in detail what do I mean by common sense, and what are those fundamental aspects that cannot be overturned. To be respectable, my thesis cannot equate common sense to any statement about society made by any untrained individual. Some people can say the most ridiculous things about society and scientists should surely refute them. But as laypeople, we also believe in some statements that are true, previous to any social science, and, moreover, self-evident, in much the same sense of the term as used in the American Declaration of Independence. That is, there are some statements about the social world around us that we simply do not prove. If someone asks us to prove them, we ought to refuse the challenge because sooner or later, any argument we give, or the very fact that we engage in arguing, must assume them. I will not offer a complete list of those truths but I will work from three of them, allowing that there may be more:

- The existence of the external world.

[1] The origins of such doctrine can be traced to different sources. Arguably, the phenomenological sociology initiated by Berger and Luckmann (1967) attempts to give scientific expression to our everyday experiences. Unlikely as it may seem, another lineage would be analytical philosophy. In fact, in this section I choose formulations that, for the most part, offer a social-scientific parallel to the ones offered by Soames (2003) in his superb account of the origins of analytical philosophy.

- The existence of the self.
- The existence of other selves, besides me.

I cannot prove these statements. Any attempt I make to suspend belief in them, while I find an argument in their defense, becomes self-defeating. I know of superb intellects that doubt some or all of them. But I claim that, as long as they follow down this path, they will arrive at spurious social-scientific doctrines (not to mention the consequences this may have for their daily lives).

These three claims may seem very mild and unlikely to be contradicted by any prevailing doctrine in the social sciences. But I will later endow them with stronger contents, albeit more controversial. What I mean by social sciences as a language for our common sense is that social sciences have no right to shatter our believe in statements like these. If a fragment of social science enters in conflict with any of them, it, and not the statements, must be rejected or modified.

3.1 Agency and the Existence of the Self

In the cosmological debates about the age of the universe, physicists have agreed on a constraint: whatever the estimate we come up with, the universe must be old enough to allow for the emergence of intelligent life-forms that ask themselves about the age of the universe. In like manner, the fundamental, self-evident truths of common sense must act as a constraint on our social theories. Any theory of society must be compatible with the existence of an external world and different selves, able of acting alone and jointly, and of asking themselves about those actions.

Unlike our naive experience of geocentrism, basic common sense is a precondition for our having experiences of the same social phenomena that social sciences try to explain. If I embrace heliocentrism, contradicting my everyday perceptions, I can carry on with the business of perceiving things, having experiences, codifying them, acting upon them, etc. But I cannot lead life as I have known it thus far if I stop believing in the existence of the external world, my self and those of others. Although there are philosophical exercises aimed at denying the existence of the external world, I do not know of anybody who in her daily life entertains such doubts. If a philosopher is crossing a street while crafting an argument that denies the existence of the external world, I wager that he will speed up to the other side if he sees a truck fast approaching without any intent of slowing down.

The most radical structuralists argue that agency is purely illusory, that social structures are so powerful in determining human actions, that there is no serious scientific sense in which we can say that individuals make choices. This violates the rule I am proposing here. The existence of selves, beings that have a mind, that feel that they have free will, that can reflect upon their choices and so on, belongs to the type of commonsensical set of propositions without which we cannot experience social phenomena. I am willing to accept a structuralist theory

that postulates that such free will is purely illusory, but only if that same theory offers an account of how could it come to happen that individuals hold such an illusion in their daily lives and *act* upon that illusion.

Structuralist doctrines fail this test when they understand self-hood as a corollary, as an accidental surplus of social phenomena. In other words, following them, it should be possible to imagine a social system identical in all respects to the current one, except that, through some twist, it does not project upon its components the notions of self-hood (e.g. free will, permanence across time, potentially infinite reflexivity, and so on). By my reckoning, such system could never be called a society.

This agent-free position is perhaps best illustrated by systems' theory, inaugurated by Niklas Luhmann. In the words of one of its leading proponents (Fuchs, 2001): "That persons have free wills explains nothing by itself. (...) [The philosophical view of free will] emerges only later, after the fact, and from an outside philosophical viewpoint. (...) Free will and agency are moral concessions, not social facts." (pg. 27) In my view, this passage is partly correct and partly false, both of them for interesting reasons. First, Fuchs is correct in pointing out that free will has no explanatory power by itself. But agency-based theories do not invoke free will as a way to improve any forecast, but rather as the raw material from which to make any forecast that can be understood as pertaining to human beings. In other words, the challenge is not only to explain why humans act the way they do, but also why, when they act, they can experience their acts as *theirs*. A witty observer of the debates in artificial intelligence once remarked that although a computer could defeat the chess world-champion, it could not feel proud about it. In that sense, Fuchs seems misguided when he considers agency as stemming out of a philosophical viewpoint; agency is anchored in our everyday life.

To be sure, we may learn a lot about society from studying such a system. Economists routinely study a fictional "walrasian economy" where the prices are determined by an auctioneer because they are confident that, at the end of the day, it is possible to translate statements about such economy into statements about real market economies. In other words, the ultimate litmus test for the plausibility of such exercise is the translation, the ability to refer the theory to something accessible through common sense. Many statements coming from structuralist theories are correct, and some even tremendously insightful. But, according to the view I have spelled out here, this can only be determined once we can connect them to our everyday experience of a social world populated by several selves.

Whether or not I am being fair to structuralism, whether or not a structuralist theory can meet all these challenges, the preceding arguments already give some plausibiliity to the fourth controversial assumption of RCT in the opening list. RCT's assumption of individuals as endowed with free will is compatible with our commonsensical views about society. In our daily life we perceive our actions as *ours*, that is, not as dictated by some deeper structure (cultural, economic, or otherwise), but instead, resulting from our own free exercise of agency. RCT respects this everyday view and elevates it to the status of a basic assumption of the theory. This is not naïveté; it is the acknowledgement that a theory of human behavior cannot blithely

dismiss the layperson's experiences of human agency. In this, RCT is implicitly assuming the commitment to articulate our common sense, rather than undermine it. As we will see later, quite often RCT wavers on this commitment.

4 Structure and the Existence of Other People's Minds.

Structuralist doctrines contain an important grain of truth: their concept of structure is compatible with fundamental experiences of our everyday life. To that extent, I am not ready to let go of structuralism and, instead, I will argue that RCT is in a good position to preserve its advantages.

In social scientific writing structures can seem very complex objects, unlike anything we are likely to encounter in our daily life. But structure-like objects are easy to come by and, in fact, are indispensable to lead our lives in society. Suppose, for instance, that I fall ill and become unable to complete this quarter's teaching of a course required in the curriculum. The department quickly finds a substitute to take over and finish the course. I have no doubt that the department will fill this vacancy with another human being, every bit as able of free will, every bit as endowed with a unique life history, as I am. This substitute will do many things differently (in all likelihood, he will be a better teacher!). The students, in turn, will react differently to him and, as a result, a lot of minutiae will change. But I cannot escape the plain, commonsensical impression that, in a way, it will be the *same* course.

Social life is structured in the sense that, although each one of us is unique, our experiences of the social world are robust to replacements across individuals. I may believe that, as result of a change of president, my country will be a better or worse place, but in many fundamental ways it will still be my country. Instead, the structures that populate social science treatises are nothing like the things I encounter in my daily life. "Political structures," "class structures," "modes of production," and so on are not things we experience directly as laypeople. Far from being an objection to structuralism, this is one of its strong points. At its best, a structuralist theory offers an account of how those "structure-like" things we encounter in daily life result from deeper structures. This procedure yields operational theories because, if we want to know how powerful a theory is, we can simply work through its implications and see if, starting from deep, hidden concepts, it can account for the things we observe every day.

Rational choice theorists often pride themselves in the privileged role they assign to agency. Whenever they intervene in the "agent-structure" debate, they locate themselves in the "agency" camp. But this move fails to do justice to one of RCT's main strengths because thanks to some of its most controversial assumptions, RCT is in a unique position to blend the "agency-oriented" and the "structure-oriented" perspectives. Fixed preferences and maximizing behavior (the second and third assumptions of the opening list) accomplish the crucial task of capturing *that part of agency conditioned by structures*.

The debates surrounding structuralism teach us that, whatever its merits, a structuralist

position is saddled with the difficulty of explaining how can structures have any effect, unless they operate through the acts of individuals. RCT provides a useful answer to this problem.

In keeping with the previously articulated views on social sciences, let's study the problem from the point of view of common sense. In our daily life, we do not experience social structures directly but through the human beings that belong to them. Structures can survive the replacement of all their members but this simply means that, however we perceive the acts of the individuals that belong to them, there is something about structures that lends predictability to that behavior. Although in the previous example, me and my substitute are unique individuals, there is something about the fact that we are teaching that makes our behavior understandable for anyone who observes us.

Social life depends to a great degree on this exchangeability across individuals. When I deal with anything that deserves to be called a structure, I need a theory of how it will work independently of the individual identity of its members. Otherwise I cannot even start interacting with it.

If I want to know what to expect from my students the moment I walk into a classroom, I need to have a theory that tells me what it is that "students-in-general" do. Individual students are unique in their responses, but I can safely predict that some things are more likely than others, by the simple fact that these are students. They are likely to sit down and (pretend to) listen what I say, speak at some particular occasions, following a set of codes (e.g. raising hands), and so on. They are unlikely to perform a satanic rite, for instance.

How do I know all this, even in the first day of class, before having met any single student? The general answer comes from what John Searle calls "the Background," (Searle, 1995) the set of conditions of the world, too large to even enumerate, that gives the world a continuity and a structure. I believe that Searle's theory of the Background is correct, but here I will focus on a more specific property of that Background: my ability to infer how individuals other than me will behave, from knowing the conditions of the external world where they operate.

Such ability is what Ferejohn and Satz (2001) call "folk psychology" and the problems it presents can be thought of as the ontological counterpart of the methodological problems with "first-" and "third-" person accounts of human action. Folk psychology offers us a theory of other people's minds, the input we need to understand the behavior of a third person. Evidently, a theory of other people's minds must have some elements in common with the theory we have about *our* minds.[2]

The thorny question is what can and what can we not extrapolate from our self-knowledge to the knowledge of others. RCT offers an answer, debatable and perhaps flawed, but a substantive answer with testable implications: the axioms of decision theory. Few defenders of RCT are willing to claim that individuals have fixed preferences and are solely motivated by utility maximization. But these assumptions capture that part of our individual behavior

[2]As Ferejohn and Satz point out, their notion of "folk psychology" owes much to the "principle of interpretive charity" advanced by many philosophers, especially Davidson (1990).

that we can recognize in others with enough precision as to base our everyday forecasts on it; the part of our behavior that is susceptible to be shaped by social structures.

Fixed preferences and utility maximization capture the stance we take toward other people in our everyday life when we interact with them within a social structure. When I walk into a bank, with the purpose of asking for a loan, I interact with several people, each of them a human being with an infinitely rich inner life, whose preferences may change and who may have many disparate, perhaps incompatible, goals. But if I want to make sense of their behavior *qua* employees of the bank, all this is besides the point. The structure of a bank is such that, in my interview with the person in charge of the loans' department, I can safely assume that this person will not approve my loan unless I meet certain requirements. Privately, this person may be appalled at having to deny loans to people who badly need them but have no means to repay them. But as far as this person is a bank employee, she will not give me the loan out of her bleeding heart. From that perspective, bank employees are exchangeable: I should expect the same response from any other individual sitting in front of me, no matter what her private feelings are. Of course, there are surprises even within structures and it could happen that I run into an employee that approves my loan, no-questions-asked. But this is an anomaly of the theory I have in mind when I walk in the bank.

In other words, when I interact with individuals within a structure, it is safe to assume that they act *as if* they had fixed preferences and utility maximization motives. Structuralist critics of RCT often claim that the assumption of fixed preferences forgets that those same preferences are the result of the structures in which individuals are embedded. This seems wrong. There is no contradiction in saying that something is the result of something else and at the same time assuming it is fixed. If the cause does not change, why should the effect do? When RCT scholars assume fixed preferences, this does not imply that they are forgetting that those preferences result from a structure. It may rather be taken to mean that, as far as the analysis in question is concerned, that structure is, for all relevant purposes, fixed (something to which the structuralists could hardly object) and that, therefore, the preferences it forms are also fixed.[3]

This is, to my mind, the ontological support for Friedman's argument. Without it, as I said before, the "as if" doctrine is of little help in evaluating and improving theories. It is not much but, RCT claims, this is all we have. RCT's folk psychology is remarkably poor: it sees individuals simply as machines endowed with powerful hill-climbing algorithms (that is what maximization comes to). There is no doubt that real-life human beings are more than that. The question is what other features (if any) of human beings belong in our basic folk psychology.

[3]This may seem at odds with Clark (1998) who, instead, tries to explain preference changes in terms of RCT. But our disagreements are rather peripheral and, instead, I believe Clark does a good job at deflating some of the objections of structuralists against RCT by showing that they are due to a misunderstanding of RCT's notion of "preferences."

5 Can There be a Complete Theory of Exchangeable Agents?

Critics of RCT are often baffled by its obtuse refusal to accept other elements in its theory of human behavior. I cannot speak for all believers in RCT but, to my mind, this refusal is largely justified, subject to the qualifications I will introduce later.

Fixed preferences and utility maximization are implausible, unrealistic assumptions about the behavior of any single person. But they serve the purpose of articulating those aspects of human agency that are susceptible to be influenced by structures. Human beings are far more complex than hill-climbing algorithms but, if the point is, as RCT implicitly and correctly assumes, to develop a third-person account of human behavior, then we ought to recognize that *no* third-person theory can capture all the complexity of a person.

In my view, RCT is not, as some of its defenders claim, a theory of human agency. It is a theory of exchangeable agents, a theory of human agents *qua* members of a structure. I mean this as a compliment. Without such a theory, we cannot understand how social structures can have tangible effects over human actions. But, by the same token, such a theory should not be confused with a complete theory of human action because human beings are far more complex than what their behavior within a structure suggests.

As critics of RCT often point out, agents can change their preferences and they can be motivated by ethical goals, rather than purely utility-maximizing ones.[4] But these aspects of human behavior cannot be theorized from a third-person perspective. Consider an example: altruistic behavior. Maybe when I go to the bank to ask for a loan, I demonstrate to the bank employee that I have no chance of being eligible and, out of the goodness of her heart, this person decides to lend me some of her own money, without even asking for a guarantee. This is surely an instance of altruistic behavior. The person who was meant to be a dour, strict evaluator of my paying capacity, has turned into my benefactor, behaving in ways incompatible with the theory I had when I walked in. But, can this change be *theorized*? No. It can be understood ex post. I can conclude that this person was outraged at the bank's policies, decided that my needs were entirely justified, that I deserved help as a fellow traveller in this valley of tears and so on. But I cannot use this as a theory. I cannot revise my theory of the behavior of bank employees for the next time I find myself in this situation. Human nature is infinitely complex, human beings are able of potentially infinite layers of reflexivity, that is, they can question their own acts, question their questioning of those acts, and so on.

In recent work, Searle (2001) has argued that no matter how many reasons an agent can muster in favor of an action, they do not constitute sufficient conditions for its choice. This is what Searle calls "the gap" in human decisions, resulting from the experience of free will that is present in every human being. As happens often to me, I agree with Searle on this point. But I part company from him where he intends to turn this into a criticism of RCT ("Classical Decision Theory" in his terminology of choice). As he points out, RCT tries to predict actions

[4]These objections have been forcefully expressed by Elster (1983) and Sen (1987).

based on reasons, an enterprise that can never have a guaranteed success. But we do this all the time in our daily lifes because we need to predict what other human beings around us will do and a rustic version of RCT, the Ferejohn-Satz's "folk psychology," is pretty much all we have. We cannot wait for a richer theory of something that, inherently, escapes any attempt at predetermination.

Critics of RCT who urge it to take into account the complexity of human agency are placing the bar at an impossibly high level. There is no theory of human agency that can capture everything there is to capture. The processes they would want to see theorized are processes that only make sense from the first-person perspective and, as such, are not subject to theorizing in the same sense as the limited agency of RCT is.

If the purported goal is to predict human actions fully, then we should acknowledge that this is a recipe for perpetual frustration: human actions are, by their very definition, contingent, unpredictable. RCT's more realistic goal is to give as close to a prediction as is possible and useful, given what we know of the objective environment where individuals choose.

6 A Reformist Proposal

In my view, the controversial assumptions of RCT represent the correct stance for the purpose they are meant to serve. We should not reject decision theory, root and branch, for the sake of an impossible standard of "realism." Instead, we can take advantage of RCT's strengths while also learning from other approaches to human behavior. In that sense, my attitude toward RCT is "reformist." To illustrate this reformist stance, I will focus on a very precise problem; I do not have an all-encompassing programmatic intention, a view of how social sciences as a whole should be practiced. I want only to show how specific social-scientific questions can be handled with the type of ecclecticism I propose.

The problem I will analyze is that of collective action. Contrary to a widely-held opinion among RCT scholars, I do not think that RCT has done a good job with this problem. To be sure, believers in RCT have given us important insights on collective action but there are lacunae too big to warrant self-complacency. In particular, RCT has failed to generate informative statements about how collective action depends on the objective context surrounding the participants.

Common sense suggests that collective action depends on its environment. Revolutions, protests, civil wars, and riots occur against the background of some objective conditions. Scholars disagree on exactly what conditions (economic, political, or otherwise) facilitate instances of collective action such as the ones mentioned. Grand theories aside, when perceptive journalists cover a civil war, they typically come up with a list of conditions that they *and the participants* agree in considering as facilitators. It may be land-tenure problems, collapse of the export economy, the recent electoral fraud, etc.

Recent scholarship on civil conflicts suggests that "grievances" cannot account for the

statistical variation across countries in the indicators of unrest (Collier, 2000). This does not refute the previous paragraph which is not a statistical claim, but rather a logical one. If I claim that behind every instance of collective action there are some objective factors, I am not saying that there must be some specific statistical correlation between collective action and sets of "independent variables" because my statement is compatible with there being many other intervening variables that affect the empirical results.

Social sciences should help us transform such common sense into a source of operational theories about specific phenomena. Instead, RCT has chosen to undermine that common sense with a parade of "paradoxes" that, supposedly, conclude that collective action should never occur and, if it did, it should not be connected with anything objective in the world. Olsonian theories claim that when collective action occurs, it is only by virtue of selective incentives that an organization provides, not by virtue of the objective circumstances that make *collective* action desirable. Analyses inspired by Schelling's notions of focal points or tipping games conclude that collective action results from a special set of mutual beliefs among the participants, but does not explain us how are those beliefs connected to the objective environment. More than forty years after the initial salvo of RCT's theory of collective action (with Olson's classic *The Logic of Collective Action*), governments keep falling under the pressure of the collective action of their citizens, outraged as they are at some specific, objective circumstances. No matter how vociferously RCT claims for itself an intellectual superiority with respect to common sense, real-life individuals will keep following their common sense and not some rarefied theory.

In a separate piece (Medina, 2004), I have presented in detail what I believe is a more coherent account of collective action, one that is compatible with our common sense and that, at the same time, clarifies RCT's strengths and weaknesses in this matter.

This is not the place to discuss at length the details of my proposal, but I will illustrate its basic points with an example. To avoid technicalities, let's focus on a very simple game with two players. Consider, for instance, the game in Figure 2. Players 1 and 2 are the only two citizens of a country and have to decide if they will overthrow a despot who makes their lives miserable. To accomplish this, they must both of them conspire against him (C). If neither of them does (both choose N), they will have to live under oppression. But, if one of them conspires and the other does not, then nothing happens to the non-conspirator whereas the conspirator will be captured and punished.

		2	
		C	N
1	C	1,1	-1,0
	N	0,-1	0,0

Table 1: Collective Action

This game has three equilibria: one pure-strategy Nash equilibrium in which the revolution

occurs (C, C), another pure-strategy equilibrium in which revolution does not occur (N, N) and a mixed-strategy Nash equilibrium where both players randomize. I will have more to say about all these equilibria but, for the time being, let's complete the enumeration of the equilibria by computing the exact formula of the mixed-strategy equilibrium.

If p_1 denotes the probability of player 1 choosing C, p_2 the probability of player 2 choosing C, and $v_i(S)$ player i's payoff from choosing strategy S, then, following classical game theory, we obtain that:

$$\begin{aligned} v_1(C) &= 2p_2 - 1 \\ v_1(N) &= 0 \\ v_2(C) &= 2p_1 - 1 \\ v_2(N) &= 0 \end{aligned}$$

The mixed strategy equilibrium is the pair p_1^*, p_2^* that solves:

$$\begin{aligned} v_1(C) &= v_1(N) \\ v_2(C) &= v_2(N) \end{aligned}$$

which in this case turns out to be $p_1^* = 1/2, p_2^* = 1/2$.

This is the moment to step back and address a few conceptual issues. First, notice that this game, which represents a small 2-person collective action problem is *not* a Prisoners' Dilemma but a coordination game. If our goal is to understand social change resulting from collective action, the Prisoners' Dilemma is the wrong template to use. It conceives of collective action as a problem of providing a public good but this is inadequate because social change is not a public good. Users of the public-goods model often overlook that public goods are things that an individual *could* provide if only he had the means to do so. But no individual can unilaterally change a government, nullify a law, introduce a new currency, etc. These outcomes are not public goods, they result from coordination. They are "social facts." In this example, a government falls the moment its citizens pull the rug from under it through their resistance. This is not something an individual can accomplish, no matter how many material resources he puts into the task.

Another important property of this game is that collective action is not a "paradox" for a theory of rationality because it *is* rational: it is one of the possible equilibria rational players could adopt. Within this model, RCT is no longer at odds with our commonsensical experience that individuals do produce social change through collective action and is, instead, a vehicle to express formally that same insight.

But if this were all, we would not need RCT. Many efforts have been devoted in RCT to "explain" collective action but this forgets that a theory must go beyond simply telling us

that something occurs; it must also tell us when and how it occurs. "Why collective action happens?" is not a good question in the same way that "Why collective action happens when and where it happens, and not in any other time and place?" is.

I claim that the method I have developed enables RCT to address this latter question. Let me go briefly through it although for brevity's sake, I will sweep many details under the rug. Suppose that upon learning that they will play this game, players 1 and 2 believe that their counterpart is likely to choose strategy C with probability $1/4$, and that these prior beliefs (henceforth, priors) become common knowledge. Then, for instance, player 1's decision problem becomes:

$$\begin{aligned} v_1(C) &= 2(1/4) - 1 \\ &= -1/2 \\ v_1(N) &= 0 \end{aligned}$$

So that N is her optimal strategy. Since the same is true for player 2 (the game is symmetric), the initial conditions described by the priors $(1/4, 1/4)$ lead the players to choose the equilibrium (N, N). The situation is reversed if the priors are $(3/4, 3/4)$. In that case, their decision problems become:

$$\begin{aligned} v_1(C) &= 2(3/4) - 1 \\ &= 1/2 \\ v_1(N) &= 0 \end{aligned}$$

In this second case, both players will find that their optimal strategy is (C, C) so that this is the resulting equilibrium.

In a sense that I will not formalize here, priors $(1/4, 1/4)$ belong to the *stability set* of equilibrium N, N while priors $(3/4, 3/4)$ belong to the stability set of C, C. Intuitively, if a prior profile belongs to the stability set of an equilibrium, this means that, if that profile is common knowledge, the players will find it optimal to play the strategies of said equilibrium. Keeping in mind that the mixed-strategy equilibrium is computed by making players indifferent among their strategies, we can realize that, once again vaguely speaking, the mixed-strategy equilibrium serves as the border separating different stability sets.

Stability sets are not part and parcel of the canon of game theory but they have an impeccable pedigree: they were formalized by Harsanyi and Selten (1988) in their work on equilibrium selection. Why are stability sets important? In my method, stability sets are important because the size of and equilibrium's stability set gives us a good estimate of the likelihood of said equilibrium. In itself this would not be very useful, but the size of a stability set is a function of the game's structural parameters, specially, the payoffs.

We can appreciate the advantages of this method if we introduce a slight modification on the previous game and study its effects over the results. Suppose that, through some change in the objective parameters, the game becomes as follows:

		2	
		C	N
1	C	2,2	-1,0
	N	0,-1	0,0

Table 2: Modified Collective Action Problem

This new game has the same pure-strategy equilibria than the previous one but its mixed-strategy equilibrium has changed. The payoff functions have become:

$$\begin{aligned} v_1(C) &= 3p_2 - 1 \\ v_1(N) &= 0 \\ v_2(C) &= 3p_1 - 1 \\ v_2(N) &= 0 \end{aligned}$$

So that the new mixed strategy equilibrium is the pair p_1', p_2' that solves:

$$\begin{aligned} v_1(C) &= v_1(N) \\ v_2(C) &= v_2(N) \end{aligned}$$

which is now: $p_1' = 1/3, p_2' = 1/3$.

If we assume, for simplicity, that all priors are equally likely, an assumption that can be easily relaxed without changing anything of substance, then we conclude that in the previous game, the stability sets of each pure-strategy equilibrium have the same size. The stability set of C, C comprises all the priors in the set $\{(p_1, p_2) : p_1 > 1/2, p_2 > 1/2\}$ while the stability set of N, N is: $\{(p_1, p_2) : p_1 < 1/2, p_2 < 1/2\}$. Since both equilibria have stability sets of the same size, then we can conclude that the likelihood of each one is 50%. That is, if we observe two players confronted with this game, we can say that they are equally likely of playing C, C or N, N.

The situation changes in the second game. Here the stability set of N, N is $\{(p_1, p_2) : p_1 < 1/3, p_2 < 1/3\}$ and that of C, C is $\{(p_1, p_2) : p_1 > 1/3, p_2 > 1/3\}$. Under the same assumption of equal likelihood for all priors, the stability set of C, C is twice as large as that of N, N. In other words, we can now conclude that the likelihood of cooperation is 2/3, compared to 1/3 of non-cooperation.

Upon closer inspection, we realize that this is all consistent with common sense. The difference between the two games is that in the second one the benefits of cooperation have increased from $(1, 1)$ to $(2, 2)$. Intuitively, this should lead us to conclude that cooperation is more likely in the second case than in the first one, a conclusion validated by the analysis of stability sets.

Apparently, this long analytical detour has done nothing else but reaffirm what we already knew. As such, one could doubt its usefulness. But the analysis of stability sets goes beyond this simple example. It can be extended to large games with N players and can help us assess with quantitative precision those conclusions that we would otherwise leave vaguely stated as qualitative insights. This is not the place to illustrate this extension of the analysis, something I have done in my book (Medina, 2004).

To appreciate the change in methodological stance that the analysis of stability sets induces, let's study carefully this 2-person country. An olsonian analysis, couched in terms of public goods, would conclude that these two players will simply free-ride on each other, thus missing the point that regime collapse is always a possibility in this country (or in any other, for that matter). The public-goods model misrepresents revolutions as outcomes that can be unilaterally produced and, in so doing, blinds us to the fact that any government, just like any other social fact, is contingent and can always be subverted, unlikely as it may be, through the absolute repudiation of those it is supposed to govern. A strategic analysis shaped by Schelling's views on tipping games and focal points will correctly recognize that an overthrow of this regime is always a possibility but will not be able to tell us how such event is impeded or fostered by objective changes in the environment. By contrast, the analysis of stability sets links the probability of overthrow to those same objective changes, giving us predictions that can be operationalized and tested. Notice that I remain deliberately vague about the exact meaning of objective changes. They may be economic, social, political, etc. Anything that is public information and that the players can mutually perceive as affecting the relative value of their actions, counts as an "objective" change. "Objective" is not the same as "physical," that would be a philistine notion of objectivity: cultural symbols can be every bit as objective as famines. The reason for this vagueness is that the analysis I have presented is not a substantive theory of social change but a method to study it. It is incumbent upon the user of the method to provide a substantive theory. As such, the method is only as good as the theory but, far from being a problem, I consider this a healthy property of this type of analysis.

A more subtle difference between the method of stability sets and the extant models of collective action pertains directly to the methodological debates outlined in the previous sections. The conventional paradigms of collective action in RCT try to reduce all of collective action to strategic calculus (as in the public-goods model) or, when they acknowledge that this is not possible (as in the schellingean tradition) remain despairingly silent about what else to do. The analysis of stability sets makes explicit that collective action is inherently uncertain because it is not only result of strategic considerations. Heeding the warnings of some of RCT's critics, this method recognizes that an analysis based solely on strategic rationality is, at best,

a third-person account of collective action and that, in any real life instance collective action involves all the complexity of agency as seen from the first-person perspective. Strategic analysis alone cannot tell us what will these two players do. If they decide to cooperate, they will do so by virtue of their personal courage, of their sense of a shared fate, of their comraderie and so on. None of these factors is strictly strategic. They all belong to that part of human agency that is not entirely shaped by external conditions. As such, game theory has no role trying to fathom them. Game theory must instead discern how their action is affected by the objective changes in the agents' environment.

Probabilistic predictions of the type produced by the method of stability sets explicitly recognize that collective action is never entirely determined by objective factors. But neither are these predictions born out of complete agnosticism. Intuitively we understand that although objective conditions never suffice to explain human behavior, their impact is not trivial and, moreover, follows certain patterns. For instance, we understand in our everyday life that increases in the benefits of collective action increase its likelihood. Debates on the causes of revolutions notwithstanding, everything else being equal, a regime is less stable, not more, if it suffers an economic collapse.

In this sense, the analysis of stability sets does to collective action problems what microeconomic theory does to the analysis of markets. Microeconomic theory does not contradict common sense. For instance, the law of supply and demand is a simple principle that can be stated without any the theoretic apparatus. But the moment it is expressed formally, it becomes a tool for empirical analysis. Likewise, in the study of collective action problems, the technique of stability sets allows us to turn into empirical hypotheses what would otherwise be untestable observations.

Apart from its contribution as a tool for empirical analysis, the method of stability sets sheds light on many of the methodological issues raised in the previous sections. In games with multiple equilibria, the principles of strategic rationality are not enough to determine one unique outcome. This has prompted some scholars in game theory to look for additional criteria that would, supposedly, narrow the set of possible predictions down to one equilibrium. But, as of yet, no set of criteria shines out as the indisputable solution. Instead, the method of stability sets invites us to accept that multiplicity of equilibria are not a pathology of the theory but rather the normal result in situations that depend crucially on mutual understandings among the players. Multiple equilibria are simply the game-theoretic counterpart to the notion that phenomena that depend on such mutual understandings are necessarily contingent; they can take many possible forms. Whereas the literature on equilibrium selection tries to induce certainty in the analysis of these situations, a certainty that has proven to be elusive, the analysis based on stability sets accepts that RCT cannot by itself yield exact predictions but that, instead, the final outcome of many social processes involves non-strategic considerations. But, just as this method recognizes the limitations of RCT, it also makes the most of its strengths. In fact, it does not simply tell us that "anything can happen," something we already knew without theory, but tries to make sense of our intuitive uncertainty by relating

it to the objective parameters of the game.

Just as important as the gains that this method allows, are the limitations it exposes. The concept of stability sets makes clear the collective action is not determined solely by strategic considerations but that, instead, its outcomes depend also on the type of mutual understandings and beliefs that the participants come to share in any given process. The forecasts made by the method of stability sets can be improved by anything that adds to our knowledge about such understandings. There is, within this method, ample room for other approaches, not guided by purely strategic analysis (e.g. ethnography), to fill important gaps in our explanations. RCT cannot hope to supersede other methods in its investigations on collective action because collective action is as much about rational choices as about communication, culture, perceptions, etc. Instead, each method must humbly offer its special strengths as part of a larger enterprise.

7 Concluding Remarks

The breathtaking assumptions of RCT have been an object of attack (and derison) ever since the inception of this scientific program. Taken in isolation, they are certainly implausible and absurd. But critics of RCT tend to overlook that such assumptions, when taken together, form a coherent response to some of the main challenges of social sciences. Explaining why, although human beings are endowed with free will, their actions follow some objectively discernible patterns, is one of the great puzzles since the birth of contemporary social sciences. Doing so in ways that articulate, rather than oppose, our everyday experiences, adds a further level of complexity to the task, one that cannot be missed without producing explanations of dubious power. RCT's axiomatic body endeavours to meet this dual challenge and, I believe, is closer to success than most alternatives.

This is not to say that RCT is perfect or even superior to other approaches. Human beings are so complex that no single theory can do them justice and a good social-scientific account of their actions must enlist the help of many different perspectives. Here I have offered an illustration of this arguing that, upon a modest reinterpretation, RCT can temper its most ambitious and implausible claims about collective action, while retaining the most valuable ones in ways that can be made compatible with what other disciplines and methods can tell us about society.

References

Berger, Peter and Thomas Luckmann. 1967. *The Social Construction of Reality: a Treatise in the Sociology of Knowledge*. Garden City, NY.: Doubleday.

Clark, William Roberts. 1998. "Agents and Structures: Two Views of Preferences, Two Views of Institutions." *International Studies Quarterly* 42(5):245–270.

Collier, Paul. 2000. "Rebellion as a Quasi-Criminal Activity." *Journal of Conflict Resolution* 44(6):838–52.

Davidson, Donald. 1990. "The Structure and Concept of Truth." *The Journal of Philosophy* 87(6):279–328.

Elster, Jon. 1983. *Sour Grapes*. Cambridge: Cambridge University Press.

Ferejohn, John and Debra Satz. 2001. "Rational Choice as Folk Psychology." Working Paper, Stanford University.

Friedman, Milton. 1955. *Essays in Positive Economics*. Chicago: University of Chicago Press.

Fuchs, Stephan. 2001. "Beyond Agency." *Sociological Theory* 19(1):24–40.

Harsanyi, John and Reinhardt Selten. 1988. *A General Theory of Equilibrium Selection*. Cambridge, MA.: MIT Press.

Medina, Luis Fernando. 2004. "Formalizing Common Sense in the Collective Action Problem." Book Manuscript, University of Chicago.

Searle, John. 1995. *The Construction of Social Reality*. New York: Free Press.

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

Sen, Amartya. 1987. *On Ethics and Economics*. Oxford: Blackwell Publishers.

Soames, Scott. 2003. *Philosophical Analysis in the Twentieth Century*. Princeton: Princeton University Press.