NATIONAL UNIVERSITY OF POLITICAL STUDIES AND PUBLIC ADMINISTRATION

DOCTORAL THESIS SUMMARY

Reasonability, reciprocity, and cooperation – A reassessment of public reason theory through an evolutionary game theoretic lens

Scientific coordinator:

Prof. Dr. Adrian Miroiu

PhD Candidate:

Oana-Alexandra Dervis

2021

TABLE OF CONTENTS

SUMMARY

In recent years we have witnessed a renewed interest in explanations that appeal to various forms of evolutionary theory in order to elucidate social and political phenomena. Two relevant examples are the neo-institutionalist and public reason traditions, both of which benefit from theoretical developments from the perspective of evolutionary game theory. Although this is not the first time such arguments have been made, the 1980s also being marked by what are now known as the "sociobiology wars", the new models proposed are much more sophisticated and promise solutions or, at least, reinterpretations of central problems for certain theoretical traditions in the political sciences.

One of the recent approaches that uses an evolutionary explanation is the one proposed by Gerald Gaus in his book "The Order of Public Reason", called the evolutionary interpretation of public reasoning. Gaus is trying to offer a new interpretation to public justification that solves the problems associated with the notion of reasonableness at the heart of the theory. This interpretation presupposes that an evolutionary mechanism of norm selection is responsible for the emergence of large-scale cooperation and proposes a number of arguments in favor of accepting the result of this selection as legitimate from a liberal perspective.

The aim of this paper was to explore the consequences of reinterpreting the theory of public reason from an evolutionary perspective and to investigate the ways in which evolutionary game theory allows us to understand interactions in the liberal-democratic public space, with an interest towards building more democratic institutions. To this end, I considered it important to answer the following four questions: What are the challenges to the project of public reason? How can game theory be used to solve some of its problems? What are the theoretical consequences of reinterpreting the theory of public reason from an evolutionary perspective? Are there other ways in which the results of evolutionary game theory can be used to support the contemporary understanding of democratic processes?

To understand the relevance of this research, the reader must be familiar with the problems associated with the notion of public justification, and in particular with the liberal project of public reason proposed by John Rawls (1993). Rawls introduces the notion of reasonableness in order to provide a solution to the coordination difficulties specific to modern pluralistic democracies.

Reasonability is presented as a feature of individuals that allows them to solve coordination problems of that rationality could not solve on its own (because rationality is oriented only towards individual interests and can operate only from the individual conception of what is good / desirable). Rawls proposes reasonable dialogue about the basic principles of society, because it is, by definition, detached from any particular conception of the good, as a solution to the problems of cooperation in modern societies, given that these problems seem to be caused by the pluralism of particular conceptions of the good (a social fact that cannot and must not be changed).

There are two major criticisms of reasonability discussed in this paper. First, reasonability presupposes the possibility of sacrificing individual short-term interests in favor of long-term collective interests. But the explanation for the emergence and viability of such a trait in human populations has long been a challenge, both for political theorists and for biologists / anthropologists. Second, reasonable behavior involves a high degree of psychological sophistication, one of the consequences of rawlsian theory being the exclusion of most real citizens from the process of public reasoning. But these two issues may become easier to address using some recent developments in game theory.

The results obtained using the evolutionary theory by theorists like John Maynard Smith (1976), George Price (1972), Robert Trivers (1971), Robert Axelrod and William Hamilton (1981), developed in the last 40 years by many other theorists, offer a theoretical solution to the first of these problems (the problem of the emergence of long-term cooperation in human communities). The prisoner's dilemma game, initially a proof of the impossibility of cooperation between rational individuals, was used by Axelrod and Hamilton to demonstrate that repeated interactions in this formula can lead, if certain conditions are met, to the emergence of cooperation. The famous equation proposed by George Price was also used to show that selection in favor of the group can, under certain conditions, be stronger than selection in favor of the individual.

The general conclusions derived from these results are that the predisposition to cooperate for the benefit of the group (and any other group-favorable feature that disadvantages the individual) can be encouraged by natural selection as long as: it increases the fitness of the group; groups in the overall population can be delimited according to the existence of this predisposition and the ability to inherit this trait is positively correlated with group membership.

The following implications of these findings are relevant when considering how they will be used in the reinterpretation of public reason. First, sustained cooperation in prisoner’s dilemma-

like situations requires the development of a mechanism by which defectors can be excluded from the group of cooperating individuals. Also, in human societies, mechanisms for excluding those who do not comply with the rules of cooperation are central to the protection of these rules, and most individuals prone to cooperation in human communities use methods to identify and punish perpetrators (displaying behaviors characterized as "*rule-following punishment*") - otherwise large-scale cooperation would not be possible. Also, an important feature of the models used is that they can also be used to explain the cultural evolution, not only biological evolution. In this sense, one of the central works in the field is *Culture and the Evolutionary Process* (Boyd, Richerson, 1985), which argues in favor of considering gene-culture co-evolution as the dominant mode of human evolution.

Starting from these discoveries, chapters II and III of the paper discuss the theoretical consequences of reinterpreting the theory of public reason from an evolutionary perspective. As already mentioned, in 2011, Gerald Gaus proposed using the results presented to reinterpret the notion of reasonableness in such a way as to provide a solution to the two issues mentioned above. To begin with, an explanation of the emergence and conditions under which cooperation can be a evolutionary stable strategy can help us understand the conditions under which reasonable behavior is possible (given that reasonableness is also a trait that involves sacrificing / bracketing individual interests in favor of public interests). Second, if it is true that all individuals living in human communities have or are encouraged to develop the ability to be *rule-following punishers*, it would seem at first glance that there is no need for sophisticated moral reasoning to solve the problems of coordination specific to pluralism. Gaus claims, based on research by Boyd and Richerson (1985; 2005), that in our evolutionary history we have developed so-called "moral emotions" that can help us solve problems related to cooperative rule-following. These emotions specific to being a "rule-following punisher" would include, among others, guilt and indignation.

On closer inspection, however, the substitution of reasonableness for "rule-following punishing" behavior has a number of less desirable characteristics from a normative perspective. For them to become apparent, we must return to the liberal principle of legitimacy, which stipulates exactly what the requirements of a mechanism for identifying legitimate moral and political norms at the level of a community should be: "*Our exercise of political power is fully proper only when it is exercised in accordance with a constitution the essentials of which all citizens as* ***free and***

equal *may reasonably be expected to endorse in the light of principles and ideals acceptable to their common human reason*." (Rawls, 1993, p.137).

Gaus (2011) refers to a series of examples and studies in psychology to show that living together has led us to develop certain moral emotional predispositions compatible with the tendency to treat our fellow citizens according to the liberal principle of legitimacy. Our behaviors in certain situations are telling: the way in which dealing with people who cannot recognize the validity of moral justifications; the way we treat children; the way we change our behavior when we understand that a rule has been broken out of ignorance, not out of malice. In all these cases we do not react emotionally or violently, we are not outraged, but sympathetic. However, these examples are, at worst, anecdotal, and at best, limited to only a fraction of the interactions in society and, in any case, involve excluding these people from the possibility of being equal citizens. It can be shown that in the recent history of mankind there are countless behaviors that involve the exclusion of an individual or a community just because they did not recognize the validity of the same rules as the majority. Also, going back to the implications of the results of the evolutionary theory of games, the same emotional moral predispositions that protect the stability of a value system can be used to protect a series of rules that can be more than questionable, because the main evolutionary mechanisms that gave rise to them are encourage conformity and the exclusion of nonconformist individuals. There are also structural inequalities assumed by some of the evolutionary models that explain the emergence of cooperation. Usually, if we admit non-atomistic interaction within the group and accept that individuals, due to their social positioning, tend to interact more with certain neighbors than with others, the distribution of strategies at equilibrium at any given time in a society will depend on the positions which various individuals occupy in that society (Alexander, 2007). In such a situation, those in key positions enjoy more freedom than other members of society. For these reasons, even if the conclusions about the emergence and stability of cooperation obtained using evolutionary game theory can describe the mechanisms that encourage cooperation and compliance in contemporary societies, they should not be used to justify arrangements that from a normative point of view are not in line with the requirements of a liberal society.

But this conclusion should not be the end of the investigation, because we must ask ourselves whether there are other ways in which the results obtained by the evolutionary theory of

games can be used to support the contemporary understanding of democratic processes. This is the investigation to which the last chapter of the paper is dedicated.

Among other things, evolutionary game theory has been used to explain the phenomenon of pluralistic ignorance (Bicchieri 2006), a phenomenon that depends on the inequalities created by the social structures mentioned above. Pluralistic ignorance describes situations in which most members of a group have a preference against the perpetuation of a certain rule or custom but feel compelled to comply with it and punish those who do not comply. This happens because they believe they are the only ones in the group who have this preference, and they themselves fear the consequences of defecting. Examples include certain religious norms in very conservative communities, some norms in adolescent groups, or some gender norms in contemporary societies. This behavior may be caused by a lack of access to information or the prominence of a minority with a favorable attitude towards that rule in key positions within the group.

The ideal of public reason presupposes the creation of an environment in which such phenomena do not influence the individuals involved in selecting the rules of a society, therefore a step towards creating more democratic institutions would involve identifying procedures that minimize their effect. In this regard, I try to investigate the idea that deliberation, under certain conditions, can minimize the effect of pluralistic ignorance, by arguing in favor of the deliberative interpretation of public reason.

Following the case studies presented in the last chapter, it can be argued that deliberation, if it takes place in an adequate environment, can minimize information asymmetry within the group or dispel misconceptions about the motivations behind the behavior of other individuals. Deliberation can also provide a framework in which attitudes change collectively, so no individual has to bear the disproportionate cost of holding the position of "trendsetter" alone. Finally, a deliberative framework creates the premises for social interaction independent of the traditionally established structures, insofar as each person can participate in the debate and his arguments are heard by all present - therefore structurally privileged members of the group can no longer exert the same influence.

The conclusions of the research are presented at the end of each chapter, the paper being organized in the form of four independent articles. However, they can be summed up by the following three general ideas. First, although evolutionary game theory can be successfully used to explain the emergence of cooperation and to investigate the conditions under which cooperative

behavior is possible, the implications of using these discoveries in a normative context are multifaceted. Second, Gerald Gaus's attempt to bring the two theoretical frameworks together should not be underestimated, but future research should pay more attention to the methodological incompatibilities between the two approaches and the implications of the findings for moral and political philosophy. Finally, although it may not have the major consequences assumed by G. Gaus, the inclusion of evolutionary models of norm selection in the research of democratic institutions can lead to useful discoveries in terms of institutional design.

References

Alexander, J. M. 2007. *The Structural Evolution of Morality*. Cambridge University Press

Axelrod, R., & W. D. Hamilton. 1981. "The evolution of cooperation". *Science* 211:1390– 1396.

Bicchieri, C. 2006. *The Grammar of Society: The Nature and Dynamics of Social Norms.* New York: Cambridge University Press.

Boyd, R. & P. J. Richerson. 1985. *Culture and the evolutionary process*. Chicago: University of Chicago Press.

Boyd, R. & P. J. Richerson. 2005. *The Origin and Evolution of Cultures (Evolution and Cognition),* Oxford University Press

Gaus, G. 2011. *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*. Cambridge University Press

Maynard Smith, J. 1976. "Group selection". *Quarterly Review of Biology* 51:277–283.

Price, G. R. 1972. "Extension of selection covariance mathematics". *Annals of Human Genetics* 39: 455–458.

Rawls, J. 1996[1993]. Political Liberalism. New York: Columbia University Press

Trivers, R. L. 1971. "The evolution of reciprocal altruism". *Quarterly Review of Biology* 46:35–57.