STUDIA METODOLOGICZNE
NR 39 • 2019, 171–193
DOI: 10.14746/sm.2019.39.7

Marek M. Kaminski

Erosion of belief in "social rationality": How game theory and social choice theory changed the understanding and modeling of social rationality*

Abstract. This article discusses how the developments in game theory and social choice theory profoundly transformed our understanding and modeling of social rationality in the social sciences due to the erosion of the concept of social optimum. I discuss the Prisoner's Dilemma and relevant examples of social situations, analyze the difficulties that arise when games are repeated, and finally, check how the main results of social choice theory influenced our understanding of the "best" social outcome.

Keywords: rationality, Prisoner's Dilemma, Arrow's Theorem, folk theorems, social optimum.

1. Introduction

One of the long-lasting byproducts of the Age of Enlightenment was the confidence in social rationality and linear progress. Bernard Mandeville (1705) firmly believed that individual vices necessarily produce social good. In Adam Smith's (1776) more refined analysis, the concept of an "invisible hand" represented the universal mechanism of aggregating multitudes of individual activities. Unconstrained markets were smart and good. Selfishly motivated actions were automatically and miraculously converted by markets into "optimal," or at least "near-optimal," social outcomes – this was an implicit

* The author is grateful for comments to Barbara Kataneksza. The support of the Center for the Study of Democracy at the University of California, Irvine is gratefully acknowledged.

meaning of "social rationality." The forces of competition and the interplay of demand and supply were inevitably pushing selfishly motivated bakers, butchers and brewers into delivering products for everybody's benefit.

The first half of 20th century brought a better understanding of the concepts of social rationality and social optimum. Pareto (1906) substituted a vague Bentham's (1780) idea of the "greatest happiness of the greatest number" with a precise definition that in a socially optimal outcome, nobody can be made better off without making somebody else worse off. These were minimal conditions since, in a typical economic setting, there are many such outcomes (called *Pareto-optimal*). Bergson (1938) introduced what later became known as Bergson-Samuelson social welfare function (swf). Any well-defined ethical or economic evaluation of social outcomes would produce its own Bergson-Samuelson swf that would assign higher numerical values to more preferred outcomes. If there were consensus in a society on its ethical and economic principles, the relevant swf would provide a perfect implementation of those principles and a measuring rod for the corresponding concept of social rationality.

The optimism of Scottish Enlightenment was attacked over the 19th and 20th centuries both by ideologically motivated socialists and communists, and by the rightist supporters of corporate state. However, the decisive blows to Smith's elegant intellectual construction were slowly hammered out by ideologically neutral developments in microeconomics that were soon represented in more formal ways in two subfields of applied mathematics, game theory and social choice theory.

An implicit – and problematic – assumption behind Smith's and similar approaches was the lack of interdependence among individual decisions, perhaps with the exception of competition among producers. To some extent, such an assumption reflected the low complexity of Enlightenment economies. Uncomplicated markets of bakers, butchers and brewers consisted of many small players who were providing simple private goods. If your neighbor bought a loaf of bread at a local bakery then, assuming no shortage, her consumption wouldn't affect your consumption in any way. In other words, in Smith's markets there were no "externalities," i.e., side costs or benefits for players other than those directly involved in an economic transaction. At that time, there was no concept of "public goods" as introduced by Samuelson

(1954) whose main characteristic was non-rivalrous access to resources such as public broadcasting or national defense, and which was later complimented by related concepts of club goods (Buchanan 1965) and common-pool resources (Ostrom 1990). For such non-private goods, consumer's utility could be strongly dependent on the actions of other consumers. For instance, if many of your neighbors used a community pool at the same time, then the crowd would make swimming utterly uncomfortable. If your neighbor switched to a public channel on her TV, you could still watch it on your TV. However, if nobody paid fees or taxes to support the public broadcaster, your public channels would stay dark. This was a more complex economic environment than Smith's markets. In such an environment, it was easily possible to obtain Pareto inferior outcomes that would be also non-optimal according to any reasonable Bergson-Samuelson swf.

Noncooperative game theory introduced especially useful tools for modeling interdependent decisions that enabled considering more complicated situations than Smith's simple markets. It taught us that humans may be locked in a variety of traps that convert their individually rational decisions into social disasters. This insight also applies to, but is not limited to, certain markets. Sometimes, unconstrained markets may lead to overexploitation or undersupply. Further developments in game theory revealed even more disturbing difficulties with the concept of social rationality. When interactions are repeated with sufficient intensity, in many games stable cooperation is in equilibrium. However, the range of equilibrium behavior may vary from total cooperation to total defection. In other words, if equilibrium is a proxy for social rationality, too many various types of behavior can be labeled as "socially rational" to make such a label useful. Finally, social choice theorists established that even if we assume the existence of a benevolent and wise decision maker willing to take into account the preferences of everybody in order to reach social optimum, the conversion of individual preferences into smart social solutions faces unsolvable difficulties. According to social choice, the very concept of "social optimum" is problematic.

The present chapter discusses how the developments in game theory and social choice theory deeply transformed our understanding and modeling of social rationality in the social sciences due to the erosion of the concept of

social optimum. I discuss the Prisoner's Dilemma and relevant examples of social situations, analyze the difficulties that appear when games are repeated, and finally examine how the main results of social choice theory affected our understanding of the "best" social outcome.

2. Game theory and tensions between individual and social rationality

The milestones in the formal development of game theory were von Neumann's (1928) proof of the existence of minimax solution in two-player zero-sum games and von Neumann and Morgenstern's (1944) comprehensive blueprint for the future development of the discipline. In the most important type of game, strategic game, two or more players independently make decisions (choose their "strategies") and the results of their individual choices result in certain payoffs for all of them. Players are interested in their own payoffs only and prefer higher payoffs to lower ones.

One of the first games that generated a lot of interest was the Prisoner's Dilemma (PD), first described by RAND mathematicians Flood and Dresher, and popularized with an eye-catching interpretation by another mathematician, Albert Tucker (Flood, 1950; Tucker 1952). This simple game describes the quintessential problem that may arise when individual decisions lead to an important social outcome.

The Tucker's story unfolds as follows. Don and Tom, two robbers, were caught by police. They are held incommunicado and expect to be charged. They care only about their own sentences; they can stay mute (cooperate with each other) or testify against their partner (defect). The prosecutor describes to Don the consequences of his decisions: if they both stay mute, there is not enough evidence for a serious charge and they are charged with a minor crime. Both inmates get one year in prison. If one of them testifies, the other one is convicted entirely on the strength of such testimony with a three-year sentence. The testifying inmate gets free. If they both testify, such testimonies are relatively less valuable and both sentences are two years. Then the prosecutor presents a symmetric scenario to Tom.

Games like the PD are most conveniently represented in a matrix form. The rows correspond to Don's (Player 1's) strategies and the columns correspond to Tom's (Player 2's) strategies. The numbers at the intersection of some row and column represent the payoffs that Don and Tom respectively obtain from playing the corresponding strategies. Since both players prefer a shorter sentence to a longer one, the numbers representing their payoffs have negative values.

		Tom	
		mute	testify
Don	Mute	–1, –1	–3, 0
	Testify	0, –3	–2, –2

Figure 1: The Prisoner's Dilemma

The dilemma is constituted by two facts.

1. First, testifying is a dominant strategy, i.e., it is always better to testify than to stay mute. This is the concept of individual rationality employed in the PD. To see this, let's consider Don's decisions assuming that he cannot influence in any way Tom's decisions. If Tom stays mute, then Don breaks free for testifying versus getting one year for staying mute; if Tom testifies, then Don gets two years for testifying and three years for staying mute. Tom's situation is analogous.

2. Second, when both inmates testify, they both get higher payoffs than when they stay mute: two years versus only one year.

As an effect, in the PD individual rationality leads to Pareto-inferior outcomes since both players could get higher payoffs under a different strategy profile. Both Don and Tom always want to testify; when they do, they are worse off than when they both stay mute.

In the generic version of PD, the strategy 'testify' is usually called 'defect' (D) and 'mute' is called 'cooperate' (C). The dilemma is between individual rationality that recommends defection and looking for social benefits from cooperation. Since individual rationality leads to inefficient outcomes, the PD provides an example of situation when an invisible hand doesn't work.

While the prison interpretation of the PD made the game attractive to present to a lay audience, it obscured its enormous importance and universal applicability. If the PD were a curiosity, its place wouldn't be at the center of political science and economics. And the applications of the PD are massive. PD was essentially the game played by the USA and the Soviet Union in the arms race. From applying PD we can learn that we do not need to assume that the opponents must be "evil": the structure of interactions forces them into defection. When we extend the basic model to many players, PD-like situations appear in global warming and other environmental problems, the formation of interest groups or the provision of all public goods and services. Even the most fundamental problem of humans forming a society versus total anarchy has arguably the structure of the PD!

The empirical problems resulting from the PD-like games motivated high-profile work in political science, economics, sociology and psychology that focused on modeling the tension between individual and social rationality. The situations involving such tensions were called collective action problems, the tragedy of the commons, social traps or social dilemmas (Olson 1965, Hardin 1968, Platt 1973, Dawes 1980). Especially important were extensions of the PD to multi-player games that retain its fundamental properties. PD can be extended in a variety of ways. In our next example, we will follow one possible path for one specific problem. Let's face the following problem of pollution in a big city:

Automobile pollution: In Los Devils, where automobile pollution is a big problem, annual benefits from clean air are estimated to be equal to $1000 per person. A pollution-control filter costs $100. We know that:
(a) the city's population is 1 mln and everybody has one car;
(b) everybody pollutes equally and the benefits are proportional to the number of filters;
(c) status quo payoff (no filters) is zero.

Every strategic game includes, as its components, players, strategies and payoffs. Let's reconstruct those components:

1. Players: 1 mln car owners;
2. Strategies: 1 (cooperate, buy filter), 0 (defect, do not buy filter);
3. Payoff of Player *i* (measured in dollars): $P_i(\mathbf{s}) = r(\mathbf{s})/1000 - 100 \times s_i$, where $r(\mathbf{s})$ is the total number of cooperators when the strategy profile is **s** and s_i is Player *i*'s strategy.

Let's consider our player's payoffs from the two available strategies when the number of initial cooperators is *r*. When Player *i* defects, his payoff is equal to *r*/1000 since he doesn't have to buy the filter. When he decides to cooperate, his payoff is equal to (*r*+1)/1000 – 100 since he must pay 100 for the filter and he becomes one more cooperator. The difference between both payoffs is 99.999. The extra benefits to a player from his own cooperation (buying the filter) are practically negligible while the cost of cooperation is high. Thus, the incentive to defect is very strong. Defection is a dominant strategy, i.e., you are always better off if you defect.

Similarly to the two-player PD, not only both players have dominant strategies but also the equilibrium in dominant strategies is inefficient. To check the second fact let's denote the strategy profile of all cooperators ALL C and all defectors ALL D:

P_i(ALL C) = 1000000/1000 – 100 = 900
P_i(ALL D) = 0

When everybody plays the individually rational strategy 'defect', everybody gets the payoff of 0. When everybody cooperates, everybody's payoff is 900. The substantial difference represents a potential gain from cooperation that is forfeited in the inefficient Nash equilibrium (see Figure Automobile Pollution). The lower line in the figure represents payoffs from cooperation; the upper line represents payoffs from defection. The fact that we do not have the continuum of players as well as the continuum of values is neglected. The dots at the ends of lines mean that we cannot have 1 mln cooperators if Player *i* defects (in such a case 999999 is the maximum) or no cooperators if Player *i* cooperates (we must have at least one cooperator).

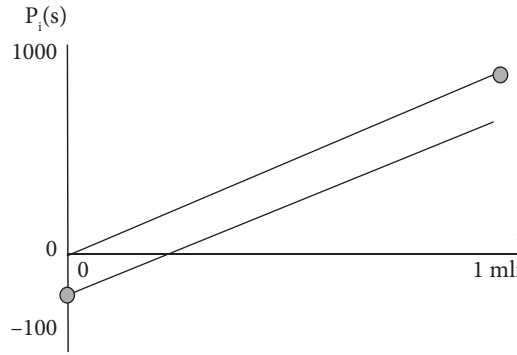

Figure 2: Automobile Pollution.

Automobile Pollution is another frightening version of the tragedy of the commons or a collective action problem. With a large number of players, achieving cooperation seems hopeless. Every player must pay the cost of cooperation but the benefits are spread thinly over a million.

From the PD perspective, it is surprising that humans so often cooperate. Cooperation may be facilitated by a variety of factors. One of them is establishing new rules of the game that essentially force players into cooperation. The rules may be imposed externally by a state or another central authority, or they may be self-imposed by interested players on themselves. Such new rules are called in political science and economics 'institutions.' The theoretical recognition of PD and similar games provided the decisive impulse for the development of modeling of institutions. One of the most profound contributions of political science was the discovery and description of the incredibly imaginative institutions that players themselves invent in order to get out of PD-like situations. In economics, increasingly larger parts of economic theory are being converted into the subfield of Industrial Organization that studies markets and other institutions that coordinate economic exchange.

In the case of Automobile Pollution, examples of game-changing institutions are easy to imagine. Sanctions may be imposed on car owners or car manufacturers that make buying a filter a dominant strategy. Products that are not environmentally safe may be banned from the market (so the strategy

of defection may be effectively removed). Informal social pressure may be imposed on carmakers to manufacture only environmentally safe products. In general, we model new institutions as modified versions of the original ones that have some strategies banned, payoffs changed by sanctions, or that were subject to more complex transformations. Cooperation is possible, but the road to cooperation is far from straight and easy.

3. Repeated games and theoretical predictive impotence

In addition to institutional change, a process facilitating cooperation is the repetition of the game. It turns out that players may routinely cooperate in PD and similar games when the game is repeated. Repeated games are typically used to model evolutionary behavior, including the evolution of social norms or repeated economic exchanges. The developments in the theory of repeated games opened its own Pandora Box of surprising effects that deeply affected game-theoretic modeling.

In a finitely repeated PD, players play PD a fixed numbers of times and the final payoff is a sum of payoffs received in all rounds. In such a game, there is no dominant strategy, i.e., against certain strategies of the opponent there may exist better strategies than ALL D (always defect). Nevertheless, the strategy profile when both players play ALL D constitutes the unique Nash equilibrium of the game, or a situation in which no player can improve his payoffs by unilaterally changing his strategy. Such a strategy profile is the one that a game theorist would be quick to predict to happen. However, surprisingly and disturbingly, players seriously deviate from total defection. The first experiment with repeated PD (in which the PD was also introduced as a game) showed this phenomenon unambiguously (Flood 1952). Out of 100 rounds, the two subjects participating in the experiment cooperated on average in 73 rounds.

The phenomenon of cooperation in finitely repeated games hasn't been explained convincingly so far. One can easily design one's own simple experiment and run it on any audience with the game of Centipede (see Figure 3).

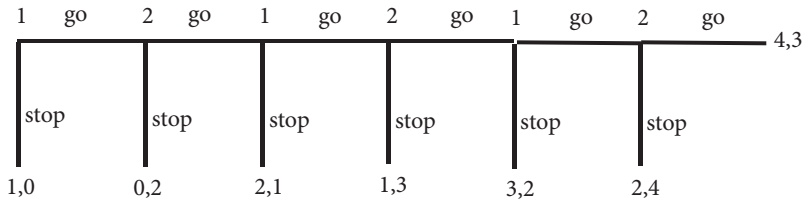

Figure 3: Centipede with three rounds.

Centipede may be considered a simpler version of the repeated PD. It can be described as follows: two players start a game with a small endowment: Player 1 starts with one dollar while Player 2 starts with nothing. They can always continue playing the game or stop it. When a player stops, both players collect the prizes that they have accumulated so far. If a player continues with the game, she must give up one dollar but the other player simultaneously receives two more dollars.

Centipede is an example of a *sequential game*. It can be solved by *backward induction* that finds all *subgame perfect equilibria*, a more restrictive concept than the Nash equilibrium. The backward reasoning is as follows: At the last stage, Player 2 wins four dollars by stopping and three dollars by going. Thus, he chooses stop. Player 1 can predict such action in her last move, and can represent her own choice as the one between receiving 3 for stopping and 2 for going (because of the predicted choice of Player 2). Thus, Player 1 stops in her last move as well. And so on: at every stage both players have incentives to stop. In the unique subgame perfect equilibrium of the game, Player 1 stops immediately, and in all possible decision situations both players would stop as well.[1]

The dilemma is quite similar to the PD. Both players receive in the subgame perfect equilibrium the payoffs of 1 and 0, respectively; if they cooperated, their payoffs would be 4 and 3, respectively. However, there is a difference between the Centipede and the one-shot PD. Experimental subjects, who are properly taught the rules of the game, tend to cooperate for a few rounds and then they stop one or two rounds before the end. Immediate stopping almost never happens!

[1] For a comprehensive evaluation and extensions of backward induction see Kaminski (2017, 2019)..

In many games, whenever interactions are repeated a finite number of times, cooperation miraculously appears. The troublesome problem for this phenomenon is that, while some explanations were offered, none is fully satisfying. Thus, the predictions made by solution concepts that we use in game theory to model rational behavior and the actual behavior in Centipede and similar games are at odds.

In infinitely repeated games, another problem emerges. Equilibria are numerous and, in many games and under reasonable assumptions, any level of cooperation may be achieved in equilibrium. In the 1970s, a number of disturbing theorems was proved that confirmed the existence of multiplicity of equilibria in repeated games. Those results were called "folk theorems" since game theorists long suspected that repeated games have similar properties.

Before we formulate one of the folk theorems, we have to introduce an infinitely repeated game created on a basis of a finite one-shot game. It is defined as follows: the players from {1, 2, …, *n*} are unchanged and non-empty strategy sets S_i (for i=1,…,n) from the original game are actions that are taken at every stage of repetition. The payoff functions are typically defined as a weighted sum of partial payoffs obtained at consecutive stages. Let *s* be a strategy profile such that at stage *k* players choose actions $s^k = (s^k_1, \ldots, s^k_n)$. Then the partial payoff of player *i* at stage *k* is equal to $P_i(s^k)$, where P_i is player *i*'s payoff function in the original game. The total payoff in the entire repeated game is equal to an infinite sum of discounted partial payoffs:

$$P_i(s) = \sum_{i=1,2,\ldots} r^{i-1} P_i(s^k)$$

where $0<r<1$ is a discount parameter. Since the discount parameter is between zero and one, and the payoffs in a finite one-shot game are limited, the series always converges to a finite number.

While versions of folk theorems can be formulated for any finite game, repeated PD received most attention from game theorists and the formulation is especially simple in this case. Let *p* denote the general payoff from mutual cooperation (corresponding to -1 in Figure 1) and *q* denote the general payoff from mutual defection (corresponding to -2).

Folk Theorem [Friedman 1971]. In the repeated PD, when the discount parameter is sufficiently high, any vector of payoffs (x,x), where $q \leq x \leq p$, may be reached in a subgame perfect equilibrium.

In other words, by varying the discount parameter, we can find a full range of equilibria: from such that both players only defect to such that both players only cooperate! The equilibrium concept, as a predictor for what will happen in a game, is under such circumstances useless.

Folk theorems provided researchers with incentives to go beyond simple analysis of equilibrium existence and to model outcomes in repeated games with the use of computer simulation. Certain results, both empirical and theoretical, suggested that cooperation may be somehow privileged against defection. Axelrod (1984) published a book that, among others, included the results of his PD 'tournaments.' He asked a number of game theorists to submit computer programs generating strategies in the repeated PD, and then played them pairwisely against each other. The simplest examples of such strategies are ALL D (Always defect) and ALL C (Always cooperate) but a strategy can be much more complex since it can take into account the information about all previous moves of both players. A strategy TFT ("tit-for-tat") submitted by Anatol Rappoport (who was one of two subjects in the first PD experiment ever played) proved especially successful. TFT starts with cooperation and then repeats the previous action of the opponent, i.e., responds with C to C and with D to D. TFT won the first tournament, i.e, it received the highest average payoff, and then, even when its success was revealed to the participants, it won the second tournament. TFT is not a dominant strategy since no such strategy exists in a repeated PD. Moreover, because of its cooperative character it cannot even win against any other strategy in a pairwise contest. Nevertheless, it seems to be very efficient in generating high average payoff across large environments consisting of many various strategies.

The privileged status of cooperative strategies in Axelrod's experiments motivated researchers to include additional variables into modeling repeated interactions. One of such variables was the 'robustness' of a strategy, i.e., informally, the minimal proportion of such strategies in the population that guarantee the existence of equilibrium. Bendor and Swistak (1997) confirmed that equilibria based on cooperative strategies, such as the TFT, are most robust.

The less efficient the strategy (in terms of cooperation), the less robust it is, i.e., frequent defectors require higher proportions to defend against intrusion. This means that frequent defectors can be destabilized by a small change in the population. This is the essence of the evolutionary advantage of cooperation.

Despite all results suggesting a special character of cooperative strategies in repeated PD, the wide range of potential equilibria remains a disturbing phenomenon.

4. Social choice theory and the problems with aggregation of preferences

The motivating questions of social choice theory differ from those of game theory. Instead of asking 'what happens when rational players make independent decisions?,' social choice enquires about the properties of various methods of making social decisions, principles of distributive justice or the existence of methods satisfying certain properties (Arrow 1951, Sen 1969, Lissowski 2013). The seeds of social choice theory were in the work of a French mathematician, social scientist, and, incidentally, one of the fathers of the Enlightenment, marquis Nicolas de Condorcet. Condorcet studied the Estates-General, the French pre-revolutionary parliament and its three main blocs of voters constituted by clergy, nobility and the others: bourgeoisie, wage-laborers and the peasantry. He noticed that voter preferences of the three groups could form a very curious pattern. When paired with majority rule, such a pattern produced paradoxes.

Let's assume that we have three voters (or homogenous groups of voters of roughly equal size) that are numbered simply 1, 2 and 3. There are three alternative policies on crime that can be implemented *x*, *y*, and *z* (e.g., spending more on police force, building more prisons and spending more on prevention). The preferences of voters in the Condorcet profile are listed below from most preferred to least preferred alternative:

1: *xyz*
2: *yzx*
3: *zxy*

Under such preferences, Players 1 and 3 prefer x to y; 1 and 2 prefer y to z; 2 and 3 prefer z to x. Thus, for every alternative, there is a majority that prefers something else to it. We have a *cycle*:

$$x \, P \, y \, P \, z \, P \, x$$

where xPy means that a majority of voters prefers x to y.

When three alternatives are at stake, voting by majority is impractical since – as demonstrated by the Condorcet Paradox – a possible outcome would be that no alternative receives majority. Thus, voting by majority over three and more alternatives is potentially indecisive. Parliamentary procedures solve this problem by making consecutive alternatives compete with each other pairwisely according to a pre-specified agenda. However, such a solution raises a question: which agenda should be used? The question is of an utmost practical importance since the relation of social preference based on pairwise comparisons using majority is, as we established, intransitive in cases of a Condorcet profile.

If some agenda setter, say, the House Speaker, can specify the order of voting, he has a considerable power in hand. If the Speaker would like to make x the winner, he could easily create an agenda supporting his wishes. Turning x into a winner requires in this case simply that x is added as the last alternative to the agenda. The winning agenda for x looks as follows:

1. Vote between y and z;
2. Vote between the winner of Session 1 and x.

According to the above agenda, denoted formally as yzx, y wins in the first round against z, and loses in the second round to x. In our example, any alternative can win provided that it is the last one on the agenda.

The *Condorcet's Paradox* showed that the naïve interpretation of voting as representing "social will" or uncovering some underlying social interest is deeply suspect. Voting can be often manipulated, and there are fundamental reasons why voting cannot be considered to be as an "objective" method of

making social decisions. Over years that followed, many other methods of voting were discovered and all of them were sooner or later found to have various troublesome or paradoxical properties. The next step in the understanding of such paradoxes was made when the contemporary discipline of social choice was created in mid-20th century by Duncan Black (1958) and Kenneth Arrow (1951). Arrow's work is especially relevant for the present article.

Arrow was a student of a great Polish logician and mathematician Alfred Tarski. Arrow's language developed for social choice theory was based on Tarski's terminology used in his popular logic textbook. Arrow claimed that he was not familiar with the Condorcet's paradox when he started his work but his approach could be considered the natural logical extension of Condorcet's problem.

Let's assume that a social decision must be made over some number of at least three social alternatives. We want to rank the alternatives from the best to the worst one, i.e., we want to order them in a *transitive* fashion (we denote this condition by T) using some social decision method F. The only information that can be used are preferences of some finite number (at least two) voters who may be also indifferent between or among some alternatives. An example of such a method would be a dictatorial rule of some voter *i* that simply says "whatever *i* prefers is also preferred by a society." Certainly, such a voting method wouldn't be acceptable for most social decisions. We are looking for such an F that wouldn't be dictatorial and perhaps satisfy a few simple properties as well. Arrow singled out the following properties:

1. Unrestricted domain (U): Our method F is defined for all possible voting profiles (configurations of voter preferences);

2. Pareto property (P): If everybody prefers *x* to *y*, then *x* must be socially preferred to *y*;

3. Non-dictatorship (D): F is non-dictatorial, i.e., there is no dictator among the voters. A voter *i* is a dictator according to F if for any *x* and *y*, whenever *i* prefers *x* to *y*, then F must prefer *x* to *y*;

4. Independence of Irrelevant Alternatives (IIA): F is based on pairwise comparisons only, i.e., if in two voting profiles the individual preferences over *x* and *y* are identical, then F must rank *x* and *y* socially the same way.

Arrow's Theorem [1951] : Conditions U, P, D, IIA listed above and the requirement of transitivity of social ranking T are inconsistent.

At least one of the five properties must be violated. For instance, majority rule applied to all alternatives violates U since for some preference profiles – such as the Condorcet's profile – there are no majority winners, and we cannot even designate the top position in the social ranking. If we modify majority by using certain pre-defined agenda and, possibly in some cases, a tie-breaker between alternatives, we can satisfy U but IIA will be violated. In general, practically all sensible voting methods violate IIA.

Arrow chose conditions that were looking "obvious" and "simple" (his original conditions were slightly different from those defined above that became later standard in the presentation of Arrow's Theorem). His point was that making social decisions is a complicated and troublesome process, far from the naivety of "social physics" and automatic rationality. 'Social rationality' cannot be modeled by a mechanical utilization of metaphors of optimization taken from physics and other natural sciences. No 'social physics' is possible. If we accept Arrow's conditions as fundamental, then every method of social decision-making must have certain fundamental deficiency or deficiencies. This pessimistic statement applies not only to voting but to all other social decisions that are based on individual preferences and result in rankings, including various policy decisions or welfare comparisons of welfare economics.

Arrow's work not only won him a Nobel Prize but also started a new discipline, where scholars uncovered a large number of similar *impossibility theorems*. Probably the most interesting and important of such results is a theorem anticipated by a philosopher Allan Gibbard (1973) and formally proved by a mathematical economist Mark Satterthwaite (1975). The Gibbard-Satterthwaite Theorem states formally what many social choice theorists suspected since Arrow: practically all voting methods are vulnerable to manipulation. Thus, this theorem extends the main point of the Condorcet's Paradox to all voting methods.

Let's assume that a voting method V is based on individual preferences of a finite number (at least two) of voters over at least three alternatives. In this case, V produces not a ranking but a single winner, i.e., a single alternative. We make the following assumptions about V:

1. Unrestricted domain (UD): V is defined for all possible voting profiles;

2. Non-dictatorship (ND): there is no voter *i* such that if *i*'s most preferred alternative is *x*, then *x* must become the winner selected by V;

3. Range constraint (RC): There are at least three alternatives that, under certain preference profiles, are selected by F as winners.

Out of the three properties, the two first ones are straightforward. The third one demands a word of explanation. The requirement of RC is in fact less demanding than the Pareto condition that appeared in Arrow's Theorem. The Pareto condition would demand in the present context that for every alternative *x*, whenever *x* is preferred unanimously to everything else, then *x* would be the winner. This would mean that every alternative must be sometimes a winner and also specify conditions when this must happen. RC demands only that we have at least three different winners, and assumes nothing about the circumstances under which this must happen.

Gibbard-Satterthwaite Theorem [1975]: If the three conditions UD, ND and RC are satisfied by V then V must be *manipulable*.

Manipulability means that there exists a preference profile such that a certain voter or voters have incentives to lie about their preferences since this misrepresentation would result in the choice of their more preferred alternative. In a less normatively loaded terminology, the type of voting resulting from misrepresenting one's preferences is called *sophisticated* or *strategic*.

Let's take a look at our example of Condorcet's Paradox and the agenda *yzx* that we used in order to make *x* the winner. Let's recall the voting profile from our example:

1: *xyz*
2: *yzx*
3: *zxy*

The agenda was: vote between *y* and *z*, then vote between the winner of the first round and *x*.

Voter 1 has certainly no incentive to vote strategically since *x* is this voter's top choice. It is left to the reader to check that also voter 3 has no such incentives. With voter 2, the situation is different. Let's assume that only voter 2 can

vote strategically. If 2 votes in the first round for *z*, his second choice, instead of his top choice *y*, such a vote would make *z* the winner of the first round. When *z* makes to the second round, it beats *x* and becomes the overall winner. Thus, by voting strategically, voter 2 was able to make the winner his second-best choice *z* instead of his worst alternative *x*. Gibbard-Satterthwaite's Theorem assures us that for all voting methods that satisfy the truly basic conditions 1-3, we can find voting profiles that are similarly manipulable.

Like Arrow's Theorem, Gibbard-Satterthwaite's Theorem generated a wave of work studying various types of manipulation or paradoxes. It turns out that problems with the agenda (as those exhibited in the Condorcet Paradox), the ubiquity of strategic voting, the manipulation via introducing fake candidates or vote trading, gerrymandering (strategic redistricting) or other electoral engineering are unavoidable aspects of politics. We cannot free ourselves from lying and manipulation simply through electing better candidates for our offices. The incentives for manipulation are present due to the nature of social decisions and there will always be politicians who will not miss the opportunity.

5. How research methodology and policymaking were affected?

The gradual dismantling of Enlightenment optimism about social rationality that happened in the middle of 20th century had substantial consequences for the methodology of research studying optimal social outcomes and the evaluation of policy outcomes. We can classify methodological reactions as *business as usual*, pessimistic *resignation*, *focus on institutions* and *fragmentation*. Below I will give examples of all types of reactions.

An example of the *business as usual* reaction is a cost-benefit analysis (CBA) that essentially employs the Bergson-Samuelson approach to evaluating policy outcomes. CBA used in public sector examines the costs and benefits associated with various policy projects. For instance, the construction of a new highway may be evaluated taking into account direct costs, environmental impact, pollution, disturbances for affected people and various types of economic benefits. In short, CBA's appropriateness stems from using monetary estimates that can be justified as well-measurable and additive across affected

parts. While problems discovered by Arrow still apply, the crucial Arrowian axiom of IIA (see Section 4) loses much of its appeal in this context.

Resignation followed the widespread pessimism about dismal prospects for collective action (Olson 1965) and the inevitability of the "tragedy of the commons" (Hardin 1968). A possible example – admittedly, a non-falsifiable hypothesis – are the problems of NATO with enforcing cooperation among its members. Olson and Zeckhauser's (1966) influential article examined the difficulties of NATO with making its members to raise defense spending to the level of percent of GDP comparable with the United States. The pessimistic conclusion was that a "large" player (United States) is doomed to be cheated by "small" players (other members of NATO). As a consequence, the United States cannot avoid making disproportionately large contributions to the common cause. Motivated by the authors' desire to show a persuasive example of a collective action problem (Zeckhauser 2015), the article quickly became a required reading for practically all graduate students of political science and public policy in the United States. Its pessimism very likely discouraged American government officials and policy analysts from pressing the allies to increase their defense spending.

Nudging NATO allies to spend more on defense by President Donald Trump provides a good example of the next reaction to the pessimism of Olson and Hardin, i.e., *institutionalism*. Trump, the first American president with no prior governmental or military experience, was probably not aware of the "impossibility" of solving the NATO's spending dilemma. He threatened with withdrawal of American troops from Europe; moving American bases to higher-spending allies; offered public shaming and cold shoulder to Chancellor Merkel of Germany; threatened with tariffs on consumer goods; made an impression that he was a "mad" decision maker. All those attempts at solving the problem were institutional in nature, i.e., he attempted to change the rules of the game in order to push the outcome in the desired direction.

In general, the explicit focus on institutions (described in more detail at the end of Section 2) has generated several Nobel Memorial Prizes in Economics over the past decades starting with Ronald Coase (1991), Douglass North (1993), and Oliver Williamson and Elinor Ostrom (2009). Especially the work of Ostrom (1990) provided brilliant theoretical and empirical arguments

against the pessimism of Olson and Hardin. While a common resource can be overexploited, Ostrom demonstrated how humans in small, local communities are surprisingly successful in changing the initial rules of game when managing fisheries, pastures, oil fields, irrigation systems or forests.

By *fragmentation* I mean the substitution of the concept of a measurable social good with various piecewise analyses that became especially popular in voting theory. Since the objective of finding the "best" voting method was unattainable, research strategies evolved towards a less ambitious goal of partial evaluation of such methods. When Arrow's work gained recognition, a popular research strategy became axiomatic analysis adopted from logic. Among its proponents, Amartya Sen received the 1998 Nobel Memorial Prize in Economics for related work while Michel Balinski and Peyton Young's results were instrumental for applying the axiomatic method to proportional representation algorithms.

William Riker (1982) famously refuted the "populist" concepts of looking for "best" politicians or implementing the "social will," and defended "liberalism" defined as merely rejecting the worst. Another fundamental research strategy in voting theory – motivated to some extent by Riker's minimalist idea of rejecting bad options – focused on the evaluation of the extent of paradoxes and problems by using computer simulation (Dougherty 2011). While no ideal voting method exists, one can try mapping the frequency of a particular problem. For instance, voting theorists confirmed by computer simulation that simple plurality method (a candidate with most votes wins) often generated paradoxical and even dangerous results in presidential elections. Thus, having a second round in such elections or using other methods was well justified.

6. Conclusion

Developments in mathematics of the second half of 20th century, especially game theory and social choice theory, demolished many of the Enlightenment's myths of a rosy society being on an auto-pilot of progress. The discovery of fundamental problems underlying the concept of social rationality gave impulse to, among others, the rise of institutional analysis that became central

for political science and economics, the use computer simulation that helps deal with many questions that are not easy to treat analytically, and the ascent of experimental methods investigating real-world decision-making.

Game theory, social choice theory and related mathematical approaches owe their modeling success to the precision of their mathematical tools and minimalistic assumptions. For instance, in strategic games described in Section 2, player identities are unimportant (they may be "bakers", "butchers" but also "states" or "voters") and strategies may come from any nonempty set; any preferences over strategy profiles are allowed. This model applies specifically to players simultaneously making independent choices (or equivalent situations), and having well-defined preferences over different outcomes – and to all such cases. We can ask questions such as "Is there any outcome such that no player wants to unilaterally change his/her strategy?" An answer will apply universally to all relevant empirical cases. Other related questions, such as dealing with coalitional opportunities or repeated play, require making additional assumptions in our model or using a different formalism. This "minimalism and precision" approach led to a development of a number of alternative modeling frameworks such as many types of noncooperative and cooperative games, repeated games, games with incomplete information, etc. This allowed for shifting the attention from providing "one explanation for all social phenomena" towards accurate matching empirical phenomena with relevant models.

The Prisoner's Dilemma and its generalizations, analyzed within the simplest framework of strategic games, destroyed the faith that markets are always effortlessly efficient and optimal. In a variety of human interactions, unconstrained decisions lead to such inefficiencies as overexploitation, undersupply or arms races. Reaching optimality via changing the rules of game, while not impossible, may be a difficult and time-consuming task.

Developments in repeated games and evolutionary game theory highlighted another troublesome aspect of social interactions. In repeated PD and other games, any level of cooperation may be sustainable in equilibrium. A troublesome conclusion may come to mind that our predictive power in such cases is impotent and "anything may happen." Surprising good news is that more cooperative strategies, such as TFT, have some evolutionary edge over less cooperative ones.

Finally, social choice theory revealed that the very existence of socially optimal states couldn't be taken for granted. Under typical conditions, no social decision rule satisfies all basic and reasonable properties that we might believe should be satisfied. Also, practically all voting rules are also vulnerable to manipulation. Whenever we make social decisions, we have to accept – consciously or not, whether we like it or not – tradeoffs between fundamental values and principles.

The pessimism about social rationality inherited with game-theoretic and social-theoretic developments had multifaceted impact on research strategies and methodology of policy making. While methods such as cost-benefit analysis remained largely unaffected, in other settings widespread pessimism labeled certain non-optimal outcomes as inevitable. Nevertheless, institutional analysis indicated that while it may be impossible to switch from an equilibrium in a game to a non-equilibrium outcome that is Pareto-superior, it is possible to change the game itself into one that would generate better outcomes. Finally, the by-product of pessimism associated with Arrow's Theorem was the emergence of the axiomatic method and computer simulation methods that instead of social optimality, investigated specific desirable properties of social decision rules.

References

Arrow, Kenneth J. 1951, 2nd ed. 1963. *Social Choice and Individual Values*. New York: Wiley.

Axelrod, Robert. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Balinski, Michel L., and H. Peyton Young. 1982. *Fair Representation: Meeting the Ideal of One Man, One Vote*. Brookings Institution Press (2nd ed., 2010).

Bendor, Jonathan, and Piotr Swistak. 1997. The Evolutionary Stability of Cooperation. *American Political Science Review* 91: 290-307.

Bentham, Jeremy. 1780. *An Introduction to the Principles of Morals and Legislation*. London: T. Payne and Son. http://www.econlib.org/library/Bentham/bnthPML.html (online reprint of original edition).

Black, Duncan. 1986 (first ed. 1958). *The Theory of Committees and Elections*: Springer.

Buchanan, James M. 1965. "An Economic Theory of Clubs." *Economica* 32 (125): 1–14.

Dawes, Robyn M. 1980. Social dilemmas. *Annual Review of Psychology* 31 (1):169-93.

Dougherty, Keith L., and Julian Edward. 2011. *The Calculus of Consent and Constitutional Design*. Studies in Public Choice. Springer New York.

Flood, Merrill M. 1952. Some Experimental Games. Research Memorandum RM-789. Santa Monica, CA: RAND Corporation.
Friedman, James W. 1971. A Non-cooperative Equilibrium for Supergames. *Review of Economic Studies* 38: 1-12.
Gibbard, A.S. 1973. Manipulation of Voting Schemes: A General Result. *Econometrica* 41: 587-602.
Hardin, Garrett. 1968. "The Tragedy of the Commons." *Science* 162: 1243–48.
Hardin, Russell. 1982. *Collective Action*. Baltimore: Johns Hopkins University Press.
Kaminski, Marek M. 2017. Backward Induction: Merits and Flaws. *Studies of Logic, Grammar and Rhetoric* 50(63): 9-24.
Kaminski, Marek M. 2019. "Generalized Backward Induction: Justification for a Folk Algorithm." Games 10.3: 34.
Lissowski, Grzegorz. 2013. *Principles of Distributive Justice.* Opladen: Barbara Budrich Publishers.
Mandeville, Bernard. 1989 (first ed. 1705). *The Fable of the Bees: Or Private Vices, Publick Benefits*: Penguin Classics.
Olson, Mancur. 1965. *The Logic of Collective Action*. Cambridge, MA: Harvard University Press.
Olson, Mancur, and Richard Zeckhauser. 1966. "An Economic Theory of Alliances." *The Review of Economics and Statistics*, 266–279.
Ostrom, Elinor. 1990. *Governing the Commons*. Cambridge university press.
Pareto, Vilfredo. 1906. *Manual of Political Economy*. Augustus M. Kelley (English translation).
Platt, John. 1973. Social Traps. *American Psychologist* 28:641-51.
Riker, William H. 1982. *Liberalism against Populism*. San Francisco: WH Freeman.
Samuelson, Paul A. 1954. "The Pure Theory of Public Expenditure." *The Review of Economics and Statistics*, 387–389.
Satterthwaite, Mark. 1975. Strategy-Proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions. *Journal of Economic Theory* 10: 187-217.
Sen, Amartya K. 1970. *Collective Choice and Social Welfare.* San Francisco: Holden-Day.
Smith, Adam. 2009 (first ed. 1776. *The Wealth of Nations*. Edited by T. Books. London.
Tucker, Albert W. 1950. A two-person dilemma. Mimeographed paper. : Stanford University.
von Neumann, John. 1928. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen* 100: 295-320.
von Neumann, John, and Oskar Morgenstern. 1944. *The Theory of Games and Economic Behviour*. Princeton: Princeton University Press.
Zeckhauser, Richard. 2015. "Mancur Olson and the Tragedy of the Unbalanced Commons." *Decisions (special issue ed. by Marek M. Kaminski)*, no. 24: 191–202.

Marek M. Kaminski
Department of Political Science and Institute for Mathematical Behavioral Sciences, University of California,
3151 Social Science Plaza, Irvine, CA 92697-5100, U.S.A.; email: marek.kaminski@uci.edu.