Peripheral Vision

Game Theory: Can it Unify the Social Sciences?

Yanis Varoufakis

Abstract

Social theorists from many different fields have hailed Game Theory as a framework that can unify the social sciences on a bedrock of mathematical reasoning, which relegates all previous attempts to provide a unifying framework for economics, political science, anthropology, organization theory etc. to social science's prehistory. After discussing in detail the five crucial theorems on which such claims are based, this paper assesses critically: (a) Game Theory's main results, and (b) the extent to which Game Theory offers a common method that can, potentially, unify the social sciences.

Keywords: Game Theory, social science, equilibrium, bargaining, evolution

Yanis Varoufakis
University of Athens, Greece

Organization Studies
29(08&09): 1255–1277
ISSN 0170–8406

Great Expectations: Game Theory and the Lure of Unifying Scientific Frameworks

An urge to grasp the far-from-obvious *essence* of things underpins all scientific inquiry. Nothing excites scientists like the discovery of links between 'bits' of knowledge which they had hitherto thought of as unrelated. They dream of stripping layer after layer of disparate, seemingly unconnected phenomena until, one day, human understanding of our world's mysterious ways is unified at some fundamental level.

Since ancient times, philosophers attempted to devise a *single* framework which would compel Nature to reveal its well-kept secrets to our enquiring eyes. The 20th century blew fresh wind into the scientists' sails as relativity theory shocked the world with the revelation that energy and matter were but two sides of the same coin. Coupled with developments in the study of the nature of light, it confirmed earlier speculation that forces and particles may not be as distinct from each other as was once thought. Optimism grew that electromagnetic, gravitational and nuclear forces will be conceived as different aspects of a single, fundamental force. The unity of natural science thus appeared closer than ever, promising some *Theory of Everything* which explains all physical phenomena within a single framework.

DOI: 10.1177/0170840608094779

It now seems that such optimism was premature. For one, the missing link between gravitation and quantum theory has proved impervious to the physicists' best efforts (see Greene 2000). However, claims for a hyper-theory have been made in recent times from unexpected quarters. Respectable social theorists have hailed Game Theory as a framework which can unify the social sciences on a bedrock of mathematical reasoning that relegates all previous attempts to provide a unifying theory for economics, political science, anthropology etc to social science's prehistory. Below I list three examples of the case for Game Theory as the great Enlightenment hope for a fully fledged, unified, genuinely scientific framework by which to unravel our ignorance regarding what makes society tick.

I begin with Elster (1982), an encyclopaedic social theorist, and author of exquisite books that transcend the simplistic boundaries of various social sciences. Though not a game theorist himself, he was among the first to articulate the claim on behalf of Game Theory: '[I]f one accepts that interaction is the essence of social life, then … game theory provides solid microfoundations for the study of social structure and social change' (Elster 1982: 457). It took some time before practising game theorists conjured up sufficient courage to jump on the bandwagon. But when they did, they did not shy away from similarly grand statements. Aumann and Hart (1992), for instance, argued that: 'Game Theory may be viewed as a sort of umbrella or "unified field" theory for the rational side of social science … [it] does not use different, ad hoc constructs … it develops methodologies that apply in principle to all interactive situations' (1982: 11) A few years later, Myerson (1999), a game theorist of some renown, compared the discovery of Game Theory's most famous result, the Nash Equilibrium (see below), with the discovery of the DNA double helix. Put together, these three statements are quite typical of the widespread belief that Game Theory, after having transformed economics over the last two decades,[1] can now pose credibly as *the* foundational *Science of Society*.

This paper has three objectives. First, to explain what all the fuss is about, by presenting Game Theory's five remarkable results, which lead so many brilliant people to think of it as the basis for a *Theory of Everything Social*. Second, to discuss some logical inconsistencies which are buried deep inside Game Theory's foundations and threaten to steal its thunder. Third, to discuss how an understanding of Game Theory and its discontents can contribute to our understanding of the social world in general and organizations in particular.

A Brief Introduction to Game Theory and its Five Main Theorems

On the Prehistory and Scope of Game Theory

Game Theory offers a comprehensive analysis of *rational behaviour under circumstances of strategic interdependence*. Suppose that you find yourself in a situation where what you want to do depends on what you think others will do. The same applies to the rest, the end result being that you have all landed in a web of predictions regarding one another's behaviour. While caught in this web,

you live in a state of what game theorists call *strategic interdependence* or, more simply, *strategic uncertainty.* The study of how rational people behave under such circumstances is Game Theory's subject matter.

As the above suggests, Game Theory is based on a particular theory of rationality: *instrumental rationality*. In short, it assumes that our *reasoning is a mere instrument in the service of pre-specified, current, and fully sovereign ends*. Jill ranks the consequences of her actions in terms of their 'utility' to *her*, and then behaves *as if* in order to maximize the (utility) rank of her actions' consequence. This is a form of utilitarianism, which neoclassical economists developed in the latter part of the 19th century. It differs drastically not only from the deontological approach of Socrates, Aristotle, John Locke or Immanuel Kant (according to whom our reason has a capacity to judge that certain actions are right or wrong irrespective of their consequences) but also from the classical utilitarianism of Jeremy Bentham and John Stuart Mill.[2]

While the above assumption about what it means to be rational is highly restrictive (and creates problems for Game Theory, as we shall see below), it does make sense in the context of real games, e.g. draughts, chess, parlour games and, more recently, poker. Playing games well is all about homing in on one's best strategy when one's objectives are well defined and in an environment where one's chances of doing well (i.e. winning) depend not only on what one does but also on what others do. Furthermore, a little mutual respect between players means that one will try to think ahead, to work out what one thinks that the others will believe that one will do. And so on, ad infinitum, with actions depending on what one thinks that others think that one thinks that … others believe one will do! At the risk overthinking matters, games are the natural habitat of *strategic interdependence*.

Unsurprisingly, the first attempts at Game Theory (just as in the case of probability theory) were motivated by games of chance and strategy between consenting players. But Game Theory was not confined to formally constituted games; it applied with equal force to social interactions in which participation was neither fun nor even voluntary. Chess, after all, was meant as an abstraction which strips the bloodshed and the boredom from the intellectual aspects of warfare. Quite naturally, any *theory* of how to play chess in order to win was of profound interest to generals, admirals and other purveyors of destruction.[3] But, at the same time, its significance stretched far beyond the masculine world of pawn, bomb and prisoner exchanges.

Suppose Tom, Dick and Harriet, three trendy young things, have been invited to a party. Each one of them wants to impress. What should they wear? Tom's choice depends on what he thinks Dick and Harriet will wear, since there is nothing more demoralizing than turning up in a frock similar to one's friend's or, perhaps even worse, in something that is totally at odds with what the others are wearing. Confusingly, Tom knows that Dick and Harriet are in the same predicament: their choices will depend crucially on their predictions of everyone's choices. So, how should one dress in this situation? Can logical analysis help? Game Theory's claim to fame is predicated on answering this question affirmatively. Its founding fathers (yes, they were all men), individuals of incredible ability, like John von Neumann and John Nash, formalized the first theorems that cemented this claim, and turned the study of games into an intellectual tour de force which is now being proclaimed the basis of a unified science of society.

To demonstrate the power of game theoretical logic it is hard to outdo the famous *Prisoner's Dilemma*. It was devised in the early 1950s by Al Tucker in a bid to impress an audience of social theorists. His story begins with two people who are picked up by police for a robbery and placed in separate cells. The police inspector visits each and says:

> 'If you deny the charges, but your partner confesses, you are facing a five-year term. But, if you confess while your mate does not, I shall intercede with the judge to suspend your sentence, on account of your assistance in bringing about a conviction. Moreover, I am prepared to put in a good word with the social security people for that pension you are after. To be frank, if you both deny the charges, I shall have to set you free due to lack of evidence. But, naturally, if you both confess, you are both going to gaol, say, for three years.'

An instrumentally rational prisoner knows that, due to the symmetry of their situation, if X is her best choice, it is also the best choice for her partner. But which is X? Of the two symmetrical outcomes 'both confess' or 'both deny', the latter is vastly superior, as it means freedom for both. However, as long as each prisoner's cares are uniquely for *her* years in prison (plus assorted private benefits like the promised pension), she is caught in a Prisoner's Dilemma, which will result in a three-year sentence! Let's see why.

Each muses: 'If my partner confesses, I am better off confessing too. And if she denies the charges, I am again better off to confess (recall the nice pension). Ergo, I shall confess *regardless of what my partner does*.' Note that their tragedy is not caused, as one may be tempted to imagine, by the fact that they cannot communicate. Even if they can talk through the cell walls, and agree to deny the charges, their individually 'best' action is still to confess! As Thomas Hobbes, circa 1651, remarked, in his famous *Leviathan*, 'covenants struck without the sword are but words'.

The Prisoner's Dilemma fascinated social theorists because it is an interaction where the individually rational choice produces a collectively self-defeating result. Each does what is in her private interest and yet the outcome is painfully sub-optimal for all. The paradoxical quality of this result helps explain part of the fascination. But the major reason for the interest is also empirical. Outcomes in social and political life are often less than we might hope, and the Prisoner's Dilemma provides one possible explanation for the frequent clash between the private and the collective interest.

As with all great theorists, Game Theory's pioneers did not 'invent' their subject; intellectuals have been pirouetting around similar issues for centuries. Sophocles' *Oedipus Rex* conveys brilliantly the power of prophecy and the problem of an infinite chain of causality between actions and beliefs.[4] Despite some recent doubts regarding its authenticity (see Meikle 1995), Aristotle (1935) analysed carefully the strategies available to slave owners for efficiently managing their estates so as to minimize any 'unproductive' reaction among slaves. Niccolo Machiavelli (1985) wrote his masterful *The Prince* with the explicit aim of coaching Florentine rulers on how to act strategically in order to maximize their control over their subjects and achieve their political goals. Thomas Hobbes' *Leviathan* (1651/1991) offered the first secular explanation of why sensible people would *want* to relinquish some of their freedoms in order to live a better life, freed equally from the fear of violent death *and* the temptation to

strike first. His argument is considered by many as an early form of Game Theory in general and of the Prisoner's Dilemma in particular.[5] J.-J. Rousseau (1762/1973) wrote cogently about the problem of organizing team production when each member of the team must choose between different degrees of commitment to a common goal whose probability of success is proportional to the effort of the least committed team member.[6] Not to be outdone by political philosophers and dramatists, Puccini's *Tosca* comes to its tragic conclusion as his heroine is caught in the jaws of a strategic trap.[7]

Edging closer to economics, David Hume (1740/1888) agonized over the chances of cooperation between neighbours when they were united by common objectives and, simultaneously, divided by doubts as to whether their cooperation would be reciprocated.[8] His pupil Adam Smith (1776/1976) conceived of his infamous *invisible hand* argument in the context of a largely strategic analysis of how public virtue may be the unintended consequence of mindless greed.[9] Less optimistic about the consequences of bourgeois demeanour, Karl Marx only had to alter Adam Smith's strategic analysis a little in order to produce a vision of capitalism as a system with an inherent propensity towards economics crisis.[10]

The examples above confirm that game-theoretical reasoning has always been with us; Game Theory simply disrobed disparate social phenomena, as well as theories about them, until their common features became visible. Without the logic of the Prisoner's Dilemma, only a person of remarkable intellect would have discerned the above similarities between Hobbes, Smith, Marx and Puccini. Today, undergraduate students are expected to draw these parallels as a matter of course. We owe this to Game Theory.

Game Theory's Five Main Theorems

Theorem 1: At least one Nash equilibrium exists in every finite game

Although John von Neumann deserves to be credited with its 'invention', the fact remains that we would not be talking about Game Theory in the 21st century had it not been for John F. Nash Jr. For it was Nash that gave Game Theory its oeuvre, with two major mathematical theorems in three articles published between 1950 and 1953. Before discussing the first of these theorems, let us familiarize ourselves with Nash's conception of a game's 'solution': what he referred to as an *equilibrium*.

Suppose that each player must choose an 'action' or 'strategy' or 'move' from a (finite) set of such choices (henceforth I shall refer to these as 'strategies'). Suppose, further, that rational thought can lead each one of them (along with us, the theorists) to a unique conclusion as to which strategy it is in her interest to choose. In this case, it is *as if* the players' thought process has converged to an equilibrium, just as surely as a rock tumbling down a hill eventually reaches an equilibrium (a 'state of rest') at the foot of the hill. Thus, a game's equilibrium is conceptualized as a set of strategies, one per player, such that the more rationally each player thinks of her 'situation', the more she tends to converge to the specific strategy in that set.

To give an example, consider the following simple *N*-person game known as the *Race-to-Zero*. *N* players are asked to write on a piece of paper (in isolation

from one another) a real number between 0 and 100 (inclusive). The player whose chosen number is nearest the *maximum choice* among all players *divided by two* wins £1m times her choice of number. (Joint winners divide the spoils.) Is there a 'solution' to this game? Is there an equilibrium towards which the players' choices will tend the more rationally they think? What number should one write down?

Nash suggests that rational players would immediately decide that it makes no sense to choose a number in excess of 50, thinking that: 'Since the largest number that can be chosen is 100, and I win if my choice is nearest to that maximum choice divided by 2, I should never choose a number above 50.' However, this thought immediately begets another, infinitely longer, thought:

> 'If I am clever enough to work this out, then the rest will also work this out too. Therefore none will select a number greater than 50, in which case I must not choose any number above 25. But if this is so, will the others not know this to be so too? And if they do, will they not restrict their choices to a maximum of 25? Then I must not go beyond 12.5.'

And so on. Asymptotically, one's optimal choice of number tends to zero just as surely as the proverbial rock rolls down a hill until, asymptotically, it hits rock-bottom. 'Choose zero' is, therefore, the game's equilibrium.

To sum up, in this case of strategic uncertainty, one's estimation of how others think is crucial. Had one's opponents been mindless machines, or monkeys, the only certainty is that one ought *not* select a number above 50. But, when playing against other rational players, and knowing it, a logical chain reaction leads each to the choice of zero. Equal winners of exactly nothing! The impetus to this ruthless outcome is none other than infinite order common belief in instrumental rationality (CBIR hereafter): As long as one believes that all others believe that one believes that all others believe … (ad infinitum) that everyone is instrumentally rational, they all choose zero.

Nash's brazen theoretical move, which allowed him to get to this unique equilibrium, was simple: he rejected all beliefs that, if held, would lead to behaviour that would have falsified these beliefs. Put differently, he admitted only beliefs that will be confirmed by the strategies that they recommend. Put differently again, Nash assumed that rational players, who recognize that their competitors are also rational, will never expect them to hold false beliefs. In the above game, it is easy to see that if one follows Nash's lead and discards all beliefs that would be contradicted by the group's choices, there is only one left:[11] the belief that each will select zero. When all players believe this, each chooses zero and the Nash equilibrium materialises.[12]

The elimination of all false beliefs does not only solve the *Race-to-Zero* (by eliminating all strategies per player except one); it also helps illuminate Adam Smith's argument, that the *invisible hand* surreptitiously eliminates the merchants' profits (just as it led the players in the *Race-to-Zero* to actions that eliminated their winnings), thus delivering the lowest possible prices for consumers. Theorists adore the confirmation of older intuition by new means, and become ecstatic when these confirmations take the form of general theorems. In this sense, Nash's great claim to fame is that, in addition to *defining* a game's equilibrium, he devised a brilliant mathematical proof that *every* finite game possesses *at least one* such equilibrium!

To see the gravity of this so-called *existence proof*, imagine that we are, indeed, persuaded by Nash that his equilibrium offers a uniquely legitimate solution to *any* game. If now every interesting social, political or economic interaction can be conceived of as some game (that is, as an interaction characterized by strategic interdependence), then Theorem 1 can be interpreted as proof that Game Theory, courtesy of Nash, has the key with which analytically to 'unlock' all types of societal phenomena.

Theorem 2: There exists a uniquely rational bargaining agreement

One may protest, rightly, that there are important interactions that do not fit into the type of game discussed above, thus weakening the claim that Nash's equilibrium can dissect *all* types of social interaction. Indeed, there are many 'games' people play in which *binding agreements* are possible *prior to action*, thus enabling players to reach decisions *jointly*, and by negotiation; as opposed to *competitively* (or, in the game theorists' own language, *non-cooperatively*). For example, organizations usually converge on action plans on the basis of collective deliberation and bargaining, and not merely through autonomous choices by isolated individuals (like those in the *Prisoner's Dilemma* or the *Race-to-Zero* games above). States, too, possess means of policing (e.g. courts, formal institutions) negotiated contracts which enable cooperative acts. Unless Nash has a 'solution' for such bargaining (or cooperative) games, Game Theory cannot pretend to offer a unified social science framework. Needless to say, he does!

To illustrate *Nash's bargaining solution*, suppose Jill and Jack are negotiating over how to share an asset of value V; an asset that they can enjoy only if they manage to reach an agreement.[13] Two conflicting forces pull their bargaining behaviour in opposite directions: the *fear of impasse* (and, therefore, to the loss of V for both) recommends a 'softer' negotiating stance, whereas the *fear of an inferior share* of V hardens their resolve.

Nash began his analysis of bargaining by stripping it to its bare bones. He assumed that, after face-to-face negotiations that last for a pre-specified period, Jill and Jack retire to separate rooms where they cool off, and, within another pre-specified period, write on a piece of paper their final claims over V: Jill claims $x_L\%$ of V and Jack $x_K\%$ of V. A 'referee' then collects their separate claims and sums them up. If $x_L\% + x_K\% <= 100\%$, they both get what they claimed (the case of agreement). If $x_L\% + x_K\% > 100\%$, neither gets anything (the case of impasse). Do *uniquely rational* claims for Jill, say $x^*_L\%$, and for Jack, say $x^*_K\%$, exist? If they do, can Game Theory predict them? it? Nash (1950) proved that, under certain conditions, the answer is affirmative on both counts.

The proof begins with a model of Jill and Jack's behaviour borrowed in its entirety from neoclassical economics (recall section 2.1); namely, the model of an instrumentally rational agent whose behaviour succeeds in bringing about the outcome that corresponds to her maximum utility, given all her current constraints, some of which are due to what other people do.[14] In this context, Jill and Jack are assumed to derive utility from their shares of V and to care *only* about the size of their *own* share. Differences in their motivation are, naturally, catered for by assuming that they may value fractions of V differently or, equivalently, they may fear impasse differently.[15]

Once Jill's and Jack's motivation has been defined, Nash (1950,1953) shows that, *given some additional behavioural assumptions* (to which I shall return in section 3 below), *there exists a uniquely rational agreement* ($x^*_L\%$, $x^*_K\%$) such that there will be neither impasse nor wastage (in short, $x^*_L\% + x^*_K\% = 100\%$). The gist of this agreement is simple: Nash predicts that Jill's share will be greater the less risk-averse she is relative to Jack. Put differently, the more Jill fears impasse (relative to Jack), the less willing she is to risk bringing it on by demanding small increases in her portion of the 'pie' and, hence, the more prone she will be to settling for a (relatively) smaller share.

In its full technical version, Nash's proposed solution to the bargaining problem predicts that Jill and Jack will settle for an agreed distribution ($x^*_L\%$, $x^*_K\%$) such that the last fraction of Jill's share (i.e. of $x^*_L\%$) yields a proportional increase in *her* utility *identical* to the proportional increase in Jack's utility caused by the last fraction of *his* share (i.e. of $x^*_K\%$). It is fairly straightforward to show that this property of the proposed agreement is equivalent to suggesting that rational negotiators will settle on *a division that maximizes the product of their utilities*.[16]

The remarkable feature of Nash's solution is his claim that it constitutes *the uniquely rational outcome of bargaining*. It is one thing to suggest some way of settling disputes and dividing pies; it is quite another to show that it is the *only* one that reason recommends. So, how did Nash prove that his proposed agreement is *the* rational one? A sketch of his proof follows.[17]

Suppose that Jack offers Jill $x_L\%$ of the pie's value V but she rejects it, demanding a higher share of, say, y_L %, and threatening Jack that, unless he relents, she will abandon the negotiations with probability p. Jill's rejection is deemed *credible* if she prefers, on average, the prospect of getting y_L % of V with probability $1 - p$ rather than $x_L\%$ of the pie with certainty. Next, let us define some agreement A to be an *equilibrium of fear agreement*, as follows: when Jill offers A to Jack, and he *credibly rejects* it in favour of some alternative division B, then Jill can *credibly reject B* (for *all B*) in favour of her original suggestion, A. Nash first proves that bargainers will *only* settle for an *equilibrium of fear agreement* and then proves that there exists only one such agreement: his proposed solution ($x^*_L\%$, $x^*_K\%$) to the bargaining problem. QED!

Noting that the above proof applies for the general case of N (> 1) bargainers, it transpires that, in a few short pages of mathematical proof, Nash seems to have derived a definitive theory of mutually beneficial agreements between rational people with contradictory interests. Let us pause for a moment to contemplate the significance of this theoretical claim. Consider the foundations of any organization, from a corporation, country club, trade union etc., to the melange of a society's legal and political institutions that determine the distribution of property and income, as well as the mechanisms for redistribution that characterize contemporary states. Are they politically *legitimate?* Can they be ethically *justified*?

Contractarians since Hobbes and Rousseau have argued that a society's institutions pass the test of justice and reason if we can *imagine* how they might have emerged naturally as part of a Grand Covenant (or Contract) between its participants, members, citizens etc. But if Nash has 'solved' the bargaining

problem by discovering a *unique solution* to it, then, at least *in principle*, Game Theory holds the key to the legitimacy (or otherwise) of organizational structures, political institutions, taxation law etc. Indeed, if there exists a uniquely rational Grand Covenant regarding the distribution of social roles between us all, it ought to be consented to by every rational citizen. The point of liberal democracy would then be to enact into a Social Contract the ... Nash solution.[18]

Theorems 1 and 2, taken together, gave game theorists the confidence to state, as Myerson (1999) does, that: 'Nash carried social science into a new world where a unified analytical structure can be found for studying all situations of conflict and cooperation' (1999: 1074). Theorems 3, 4 and 5 reinforced this claim significantly.

Theorem 3: Extension of Theorems 1 and 2 to dynamic settings

Real life requires real time. However, Nash took the sequence of our actions off the agenda by creating a real-time vacuum in which he forged his solution-concepts (note how, in the preceding games, players choose once and simultaneously). Would Nash's 'solutions' survive an infusion of real time? It was left to his intellectual ancestors, especially John Harsanyi and Reinhart Selten, to prove that they may well do. To illustrate, consider the following simple sequential (or dynamic) *7-coin game*.

On a table there are seven rare gold coins of tremendous value. Two players, Ann and Bill (A and B hereafter), are given an opportunity to collect them as long as they abide by the following simple rules: A is invited to approach the table and take either one or two of the seven coins. If she takes two, the game ends. If, on the other hand, she collects a single coin, then B gets to approach the table and take either one or two coins. Again, if B takes two coins, the game ends. But if he collects only one, A gets to revisit the table. And so on, till either a player collects two coins at the same visit or no coin is left on the table.

The reader will immediately notice that this is no more than a dynamic, multi-stage version of the *Prisoner's Dilemma*: taking one coin is the equivalent to the cooperative strategy (of not confessing to the police) while taking two coins is tantamount to defection from the cooperative outcome. Moreover, while it is in the interest of both players that neither collects two coins, as long as there are more than two coins left on the table, each has an incentive to 'cheat' by doing precisely that. To see this, game theorists invite us to enter into A's shoes, at the outset, and reason as follows:

> 'In the first round of the game, with 7 coins on the table, what should I do on my first visit? Take one coin (thus giving B a chance to play) or take 2 and "kill" the game instantly? It all depends on what I think B will do in the second round if I take only one now. If I predict he will take 2, that will leave me with a single coin and, therefore, I might as well take 2 now. If, on the other hand, I think he has good reasons to take only one, thus giving me a chance to revisit the table in the third round, then it may be a good idea for me to collect only one now.'

It sounds as if A has cause to think in terms of mutual (or *tit-for-tat*) cooperation. If A thinks that B will cooperate by taking only one coin when there are six left on the table, she is willing to cooperate at the outset (by taking only one coin). However, a combination of Nash's logic and the so-called logic of *backward*

induction wrecks the prospects of such cooperation. The reason is simple: in trying to work out what will happen later on, when there are fewer coins on the table, she reaches the sad conclusion that neither of them will be able to bind themselves to the strategy of collecting a single coin *at any of their visits*. For instance, she knows that if the game's fifth round is ever reached, there will be only two unclaimed coins and it will be her turn to play. Obviously, she will collect both! But will not B have predicted this in the previous (fourth) round, when he visits the table, with three remaining coins?

Of course he will. In a bid to pre-empt her next move, he will thus remove two of the three remaining coins in the game's foruth round (thus ending the game there and then). Having worked this out from the outset, A concludes that if she ever gets a chance to return to the table in the game's third round, with four coins left, it will be *she* who pre-empts B's pre-emption (by removing two of the remaining four coins). But, then again, B will have worked that out, too, in the second round, and will not give A any chance to get to the third. Having foreseen all this at the very beginning, A kills the game in its first round by collecting two coins immediately, thus wasting the golden opportunity given to both A and B to collect, among themselves, the complete collection of seven priceless coins.

In conclusion, games in which players move in real time, are 'solved' by Game Theory by the above combination of Nash's logic and backward induction. Starting from the game's end, rational players derive their optimal strategies for each round on the assumption that each player will act in a manner that is the best reply to the moves of their opponents once all potential future moves have been assessed.

Theorem 4: Extension of Theorems 1,2 and 3 to risky settings

Contemporary market societies are characterized by unprecedented inequality. Defenders of capitalism traditionally argue that the fabulous wealth of the few is legitimated by the risks they had to take. Profit is, therefore, portrayed as the just reward for making risky choices in an ocean of uncertainty. The problem with the type of Game Theory so far encountered is that there is really not very much uncertainty: players know the rules and share perfect information over the others' objectives, thought processes etc. Unless Game Theory's results extend to environments in which players are genuinely uncertain, its grand claims will seem hollow.

It was John Harsanyi who offered the decisive theoretical proofs that Game Theory can incorporate risk with aplomb. His method involved two steps. First, whenever faced with an opponent whose character is somewhat opaque, we are encouraged to think of her as one randomly drawn from a population of N possible opponents. As long as we can assume that one knows all the *possible* types of one's opponents, and assign a subjective probability to each of these types being *the* actual opponent, then it is *as if* one is playing a game against N opponents, each with her own character (i.e. payoffs) and likelihood of being one's actual opponent. Second, find the Nash equilibrium (or bargaining solution) to this $N + 1$ game by assigning to the payoffs of each of one's N potential opponents a 'weight' reflecting this likelihood. The resulting equilibrium is discerned

by employing Bayes' rule,[19] and has come to be known as the *Bayesian Nash equilibrium*.

In this sense, when bargaining with some person whose motives escape you, it is *as if* you are bargaining not with one but with an army of N negotiators. Each of these N 'characters' represents *one* possible motivation of your single opponent. And if your opponent is similarly uninformed about your motivation, she is also (from her perspective) bargaining with M versions of yourself. So, what in reality is a two-person negotiation becomes a republic of $N + M$ *potential* characters. The gravity of the motivation of each one of the $N + M$ characters on the final solution/contract is proportional to the likelihood that *that* character is the true character of one of the two bargainers.

Theorem 5: Evolutionary equilibria are Nash equilibria

Sceptics have often cast doubt on Game Theory's capacity to illuminate social phenomena due to its dependence on 'too much rationality'. They argue that society is inhabited by people who laugh and cry, often act spontaneously against their better judgements, follow one another like sheep, are neglectful of their interests, even populate the psychoanalysts' waiting rooms as a result of self-loathing. A theory that presupposes self-interested agents with infinite computational powers is, surely, incapable of explaining any of the above, the repercussion being that Game Theory's claims to offer a foundation for a general science of society should be taken with a large pinch of salt.

Game Theory's greatest forte is its capacity to turn criticisms into a source of strength. The criticism in the previous paragraph is a good case in point. Suppose we have a large population of players who are totally devoid of rationality. They interact repeatedly, each time with fresh partners/opponents, and have little or no idea of how their 'payoffs' are being determined. In fact, let us assume that, initially, they act randomly and thereafter mimic the behaviour that seems to be relatively more successful. From time to time, some player adopts a weird behaviour either because she is experimenting with an alternative strategy or because of an error. Note that the above typifies an evolutionary process: There is an *adaptation mechanism* (players mimicking the 'successful' behaviour) and a *mutation mechanism* (which throws up mutant behaviours with small probability in each round of a game).

The above context could not be more different to that which Nash had in mind. Rationality being conspicuous by its absence, there is, indeed, no model for players in that context. Since actors have no notion of the interaction in which they are involved, it is neither interesting nor possible to model their decision making. This is not surprising. Evolutionary biologists like John Maynard Smith and Simon Price developed such models in the context of studying the behaviour of ants and birds: a context that renders all talk of commonly known rationality and long-term planned strategies superfluous. And, yet, the findings are startling from the perspective of Game Theory.

Take any of the games that we discussed above, the *Prisoner's Dilemma*, the *Race-to-Zero* or the *7-coin game*. In our analysis, we had assumed that players were not only rational but that they were forming their strategies on the assumption of an infinite order of common belief in each other's rationality.

Let us now adopt the evolutionary perspective described above; that is, assume that our players are complete idiots: automata mimicking the most successful behaviour and occasionally behaving unpredictably with no rhyme or reason. What will happen then? It is easy to show that the equilibrium of this evolutionary process will always coincide with ... a Nash equilibrium![20]

This is a truly remarkable result. Game Theory's main tool for dissecting social interactions, the Nash equilibrium, seems to be making sense in two violently different settings: (a) in a world of hyper-rational players who act on the presumption that everyone is like them, *as well as* (b) in a world of idiots who mindlessly grapple towards a behavioural equilibrium on the basis of mimicry and random errors/deviations. If Game Theory, as it transpires, has the concepts with which to get to the bottom of both types of societal formations, then its claim to offer a unified framework for the social sciences may not be without foundation.

Game Theory's Achilles Heels

The indisputable appeal of its five magnificent theorems notwithstanding, Game Theory features two Achilles heels that place its grand claims in jeopardy: First, there is what I shall term *Radical Indeterminacy*, a condition caused by a proliferation of 'equilibrium solutions'. While Nash proved the existence of at least one equilibrium for each conceivable interaction, the number of such equilibria tends to explode, the more interesting the interaction under study.[21] Thus, Game Theory loses explanatory power, as almost any outcome can be depicted as the outcome of rational play; but a theory that rationalizes everything explains, in the end, very little.

Second, there is the problem I describe as the *Rational Deviance from Equilibrium*. Even in games that Game Theory claims to have 'solved' by identifying a unique equilibrium or solution (e.g. in the *7-coin game* or the *Bargaining Problem*), there are serious grounds on which to dispute its conclusions. Take for instance the *7-coin game*. The equilibrium 'solution' is that the first mover (player A in our narrative) collects two coins in the first round and the game ends there and then. If players are rational, and this fact is common knowledge, player B ought to expect with complete certainty that he will not get a chance to visit the table, since A will stop the game in its tracks by collecting two coins in the first round. Well, what if A collects only one? What will B think?

This is a tough question. If a zero-probability event were to occur in Nature, our model of the world would collapse (e.g. if the Sun were to refuse to rise tomorrow morning). Thankfully, in a social context, a host of milder alternatives are available. In this game, B may well think that his belief in A's rationality might have been misplaced. Having observed that she collected one coin in the first round, against the edicts of Game Theory, B may come to question A's rationality. Suppose that, in the second round, which occurred only because of her 'error' in the first round, he has revised the probability that 'A is a person who irrationally collects 1 coin' from $p = 0$ to some positive value $p = p^*$. Suddenly, it is not immediately obvious that he must collect a single coin himself: For if,

p^* is high enough, he may be better off collecting one coin in the second round in the hope that she will do the same in the third round. B's objective now becomes to give A the opportunity to keep collecting one coin until he steps in, in some later round, and collects two coins.

The remarkable observation, here, is that the above thoughts may persuade even a rational A to violate Game Theory's advice and collect only one coin in the first round. For if she has reasons to think that by so doing she will cause B to revise downwards her reputation as a rational player from 1 to $1 - p^*$, thus making her ripe for a decision to collect only one coin in the second round, she may well 'invest' in a reputation for irrationality so as to increase her coin tally beyond the measly two coins that Game Theory is advising her to settle for! Of course, nothing can guarantee the success of this bluff. Indeed, B may recognize A's initial decision to collect one coin as a bluff. In this case, he will *not* revise downward his subjective probability estimate that she is rational (p will be left at zero) and collect two coins in the second round (as Game Theory advises him) without a second thought. However, this behaviour *cannot* be uniquely rational either. For if it were, it would be clear to A that bluffs *never* work. But if that were common knowledge, no rational A would ever collect a single coin in the first round, in which case any observation of an A collecting a single coin in the first round ought to mean that she is irrational, and thus B (in that case) would always collect two coins in the second round. But were this true, a rational A would *always* benefit from bluffing. A contradiction!

In a nutshell, it may be rational to act irrationally (as all good bargainers and poker players know) while, at the same time, such investments in a dodgy reputation never work in equilibrium. This is the same, however, as to dispute Nash's belief that to isolate a game's equilibrium is to solve it. Rational action in a social context may deviate systematically (but crucially unpredictably) from any equilibrium that Game Theory comes up with, however clever its derivation.

A similar criticism is in order in relation to Game Theory's conclusion regarding bargaining. Clearly, much hinges on the alleged *uniqueness* of Nash's bargaining solution. If it can be authenticated, game theorists will have gained privileged access to the idea of both the efficient operation of markets and of the Good Society. However, a problem similar to that in the previous paragraph throws a spanner in the theoretical works. Suppose that the uniquely rational bargaining solution (e.g. that offered by Nash) to a bargain involving N negotiators instructs negotiator i to accept share s_i of the pie. Suppose further that, while all N players labour under a common belief in each other's rationality, negotiator i startles all by refusing to accept share s_i, insisting on $s_i + \delta s_i$ instead. How will the rest react to such recalcitrance?

The point here is that there is nothing in the theory (and there can never be anything in it) that offers them guidance on this issue. Thus, the remaining $N - 1$ negotiators must devise responses that the theory, by its own construction, cannot provide. But if that is so, there is no reason for negotiator i to believe that it is *necessarily* irrational to deviate from the theory's advice and settle for s_i. And if that is the case regarding negotiator i, it must also be true for all N negotiators. In short, no bargaining theory can unearth the uniquely rational agreement among clever negotiators.

In summary, this section has identified two grave problems embedded in Game Theory's foundations: The first one is that social life, especially when real time is allowed into the analysis, generates a plethora of equilibria among which Game Theory is at a loss to identify the ones that are more likely to occur. The second one is that, even when the theory can identify a unique solution or equilibrium, human reasoning can lead to behaviours that violate the unique 'solution'. We do this not because, as players, we are not rational enough to emulate Game Theory's solution but, on the contrary, because we are capable of higher forms of reasoning than Game Theory is prepared to acknowledge. This is crucial. Most critics of Game Theory accuse it of assuming too much rationality. Here, I advance the criticism that Game Theory *underestimates* the subtleties and subversive capacity of human reasoning. In so doing, its predictive power regarding social phenomena suffers.

There is an important link between this criticism and Game Theory's celebrated treatment of *uncertainty*. The examples of bluffing, above, made clear that rational people have a capacity to shroud themselves in a cloak of deception that injects uncertainty in the minds of their opponent and, thus, helps them achieve their objectives better. This is a kind of uncertainty, however, that Game Theory cannot handle. In the exposition of Theorem 4 (see section 2.2), we saw how game theorists model uncertainty by imagining that, while one may not know the character of her opponent, she knows the complete distribution of her *potential* opponents: a distribution commonly known among all players. The important point about deviant behaviour and rational bluffs is that they engender a deeper sense of uncertainty: one where not only do you *not* know the precise type of your opponent but you are equally ignorant of the *probability* with which she is of one or the other type, as well. Unfortunately, mathematical derivations of solutions under this type of *genuine uncertainty* are impossible without devaluing the phenomenon under study.

Lastly, a point on the relationship between *Rational Deviations from Equilibrium* and the evolutionary turn of Game Theory (recall Theorem 5). We saw how Nash's method was reinforced by the finding that all evolutionary equilibria are also Nash equilibria. However, that proof was based on an assumption that is questionable in social science, although quite believable in biology. While I have no qualms with the idea that modelling mutations as random and independent events does not jeopardize biology's predictive power (viz. the evolution of genes, phenotypes etc.), I strongly doubt whether this is an adequate assumption in the social sciences. In human organizations, be they universities, corporations, book clubs or Parliament itself, the mechanism that generates variety (that is, 'mutant' or deviant behaviour that strays from established norms) is never *statistically independent* of the adaptation mechanism that selects among behaviours. Behavioural deviance or 'mutations' within human communities have the habit of becoming highly co-integrated with collective behaviour, as people with common interests seek, often through dialogue, to coordinate their subversive acts against conventions that have either been established or are in the process of being so. The presumption that human society's mutation mechanism is 'apolitical' is one of several reasons why the evolutionary turn of Game Theory misses a great deal of that which matters in the evolution, or history, of human societies.

4. Discussion and Conclusions

Undoubtedly, Game Theory offers a unified account of all things social. Based on Nash's theorems and their extensions (recall section 2), it suggests fascinating analyses of anything: from the marketing strategies of ice cream sellers on Californian beaches to Rousseau's Social Contract; from art auctions in London to the history of Latin American rebellions; from the evolution of business norms in the United States to the role of taboo and ritual in Papua New Guinea. As each day goes by, the literature that extends Game Theory to every conceivable aspect of human (and some non-human) activity proliferates. Nevertheless, the sceptic has plenty of reasons to doubt whether Game Theory is an adequate foundation for a unified science of society. There are two types of doubt here: concerns with grand theoretical claims in general *and* doubts regarding Game Theory's particular problems (some of which I discussed in the previous section).

On the first type of doubt, it suffices to point out that it is one thing to offer a unified approach to all things social, but quite another to unify the social sciences. An effective unification requires more than an ability to engage with all social phenomena; it requires that the resulting analysis retains the rich perspective brought to the subject by each of the, hitherto separate, strands of social science. Granted, for instance, that Game Theory can suggest an internally consistent explanation of rituals in some remote African village or the norms of certain corporations, what do we miss out when traditional anthropological notions are jettisoned because they are at odds with instrumental rationality? Critics of grand meta-narratives tirelessly point out that the urge to unify disparate theoretical perspectives is underpinned more often by the need of the theorists to amass greater power for themselves (e.g. in the corridors of the great universities) than by an honest craving for enlightenment.

Besides the general scepticism with grand theoretical claims, Game Theory's audacious assertion invites criticism specific to: (a) its *assumptions* (for instance, that agents are instrumentally motivated and that they have common knowledge of this narrow form of rationality); (b) its questionable *inferences* drawn from these assumptions (as when it is *assumed* that common knowledge rationality delivers consistently aligned behaviours and beliefs; see the *7-coin game* in section 2); and (c) the failure (even once the controversial assumptions and the inferences are in place) to generate determinate predictions of what 'rational' agents would, or should, do in important social interactions (i.e. *Radical Indeterminacy,* see section 3).

The reader would be excused, at this juncture, for asking: 'If Game Theory cannot be relied upon to explain *all* social phenomena, what kinds of behaviour *can* it explain adequately? What are its successes and which are its failures?' While it is possible to answer by pointing to socio-economic situations in which Game Theory's predictions have a greater chance of confirmation (e.g. auctions), a broader point is in order here. At least epistemologically, the social world differs radically from Nature, where it does make sense, for example, to use simple Euclidian geometry when designing houses, straightforward Newtonian physics when building jet engines, and only turn to the mind-boggling complexity of Einstein and Hawking for phenomena that happen far away

and at very high speeds (or phenomena so far hidden in the microcosm that our best microscopes cannot detect them).

In sharp contrast, social phenomena cannot be classified so neatly between 'simpler' and more 'complex' realms, before deciding to deploy different approaches, like Game Theory, in some but not in others. For even the most mundane human act is laden with social meanings. A simple purchase at a department store, the casting of a vote, the choice of an advertising strategy, the waving of a flag — indeed, any situation featuring human minds — our acts generate meanings that demand our full analytical powers in every realm. In short, in humanist science there is no equivalent to a realm in which Eucledian geometry will suffice.

To illustrate, let us consider once more the simple Prisoner's Dilemma of section 2.1. Game theorists are convinced that it is a truth of logic that instrumentally rational players will fail to cooperate regardless of whether the game is played between governments at the World Trade Organization or by commuters trying to squeeze into a train carriage. Philosophers, like Martin Hollis, who used Game Theory extensively in their work due to its unquestionable pedagogical value, argue that this conclusion is scandalous.[22] Hollis's point is that there is no neat separation of (i) the way that we conceive an interaction, from (ii) our reasoning that leads to a conclusion on how we ought to act within that interaction.

Take, for example, the utility that Jill gets from the mutual confession outcome in the Prisoner's Dilemma. Game theorists assume that all that matters in determining Jill's utility is the outcome itself: that they both confess and will now do time. Theorists like Hollis protest that Jill's utility cannot be independent of her perception of Jack's *reasons* for defecting. For instance, compare cases (a) and (b) below:

Jill confesses, expecting Jack to confess too

(a) because (she thinks) he fears that she will confess.

(b) because (she thinks) he hopes that she will not confess (and, therefore, that he will get out scot-free, with the added bonus thrown in by the police, while she languishes in gaol).

Even though the outcome is identical (they both confess), it may very well be the case that Jill gets more utility under (b) than under (a), since (b) also gives her the satisfaction of having prevented a mean Jack from benefiting from his meanness.[23] If this is so, notice the point's analytical significance: all of a sudden, the *same* outcome (mutual confession) gives Jill different utility depending on her beliefs about Jack's beliefs. Similar arguments apply to all of the game's feasible outcomes. For example, Jill may gain utility from the thought that, if Jack were to expect her to deny the charges, he would deny them too, thus forfeiting the informer's prize for her sake. This raises the possibility that her utility from the mutual denial outcome rises above her utility from getting out of gaol at Jack's expense. But if this is so, suddenly, cooperation becomes possible and the Prisoner's Dilemma structure may be *overcome* by the players' own instrumental reasoning.[24]

So, it seems that human reasoning can indeed subvert Game Theory's firmest conclusions by dissolving the assumption that the players' motives can be specified fully *prior* to the agents' strategic thinking about the interaction. This conclusion applies equally among highly skilled negotiators, government ministers and petty thieves: social complexity of the highest order emerges in the most mundane, just as in the grandest of, contexts. Put differently, there is no such thing as a mundane social interaction! We are not at liberty, therefore, to say: 'Here is a simple enough interaction; let's apply simple Game Theory to it.' As long as one human is present, the phenomenon at hand is arguably as complex as it can get, and a game's rules do not only constrain agents; in a manner that resonates with the late Wittgenstein, rules also play a constitutive role in that they help agents (re)define the game.

In recent years, three major developments have helped increase our fascination with Game Theory: its association with *evolutionary theory*,[25] its application to the *design of large-scale auctions* (e.g. the auctions of 3G and 4G telephony spectrum),[26] and some clever *laboratory experiments*.[27] However, while all three have added to Game Theory's kudos, at the same time they have combined nicely to reinforce the criticism in the last few paragraphs above: the point that *action* and the *social structure* in which it occurs are linked complexly, *dialectically* as philosophers might say. But this is a point that game theorists are loath to admit. Traditionally, they have shied away from this relationship. Instead, they invest greatly into the instrumental account of the social world, which begins with pre-manufactured interactions that are then 'peopled' by agents who act mechanistically, with given motives and reasons, yielding (through their choices) social outcomes, market equilibria, organizational structures, political institutions etc. The explanatory traffic is singularly one-way: from individual action to social and organizational structure.

However, all theoretical and empirical evidence points to a reality closer to the dialectical or Wittgensteinian view of the penultimate paragraph above; one in which agents do not only form organizations and shape society but where, at the same time, social organization is being incessantly embedded in individual agency. Game theorists treat organizations as the crystallization of individual action. Perhaps the time has come also to view individuals as complex, evolving organizations. If the great variety of contemporary social science is converging on a single view, it is the view of real people who appear to be more complexly motivated than Game Theory's instrumental model allows for. Moreover, a part of that greater complexity comes not from 'irrationality' but from their social location. As long as Game Theory turns a blind eye to this, its offerings lose their gloss the more we subject them to rational scrutiny.

Nevertheless, the above criticism is not meant as a negative conclusion vis-à-vis Game Theory's contribution. Quite the contrary, I submit that critical engagement with the five theorems of section 2 is the ideal sounding board for any challenge to the type of methodological individualism which has had a free rein in the development of Game Theory, in particular, and the economic approach to human behaviour, in general. The problems of *Radical Indeterminacy* and *Rational Deviance from Equilibrium* (see section 3) need to be addressed by any science of society. Even if the claim to have unified the social sciences rings hollow in some ears, Game Theory has made a significant

contribution by inciting a fresh dialogue between the social sciences. Perhaps the most helpful conclusion of this dialogue is the thought that adequate social theory requires either that a greater organizational complexity (and its social dimension) be coherently incorporated in an individualistic framework, or that the methodological foundations of modern social science shift away from individualism. Game Theory, in this sense, forces us to be ambitious in ways that game theorists have possibly not imagined.

Notes

1 Soon after its birth, Game Theory developed a symbiotic relationship with economics. Indeed, since 1994 six game theorists have been awarded the Bank of Sweden (Nobel) Prize in Economics. In this day and age, most respectable Economics Departments are being populated by academics with, at the very least, a strong background in Game Theory.

2 By treating all politics as reducible to the instrumental acts of atomized individuals, Game Theory confines itself to a particularly narrow form of *liberal-cum-methodological individualism*. Institutions, ideologies, norms etc. are explained in terms of behaviour driven solely by Jill's and Jack's 'utility' ranks, which are (a) bleached of all moral and social psychology, (b) interpersonally incommensurable, and therefore (c) incapable of suggesting whether it serves the *Common Good* that a certain prize or burden be assigned to Jack or to Jill. While well placed fully to explore the conceptual limits of any theory that models society as a contested terrain on which atomistic agents act, Game Theory's insights are limited by the assumption that agents have no capacity to submit their own (and others') preferences (over outcomes) to rational scrutiny; a capacity that some (e.g. Rousseau, Locke) say distinguishes a *ζώον πολιτικόν* (political animal) from the brute.

3 Indeed the first flourish of Game Theory in the United States during the 1950s was financed to a large extent by the Rand Corporation, the Pentagon and the US Navy in a bid to design the best strategic plan for thermonuclear warfare. In fact, a caricature of John von Neumann, Game Theory's founding father, appeared in Stanley Kubrick's anti-war comedy *Dr Strangelove*. Peter Sellers, appearing in multiple roles, portrayed von Neumann as the half-crazed, 'bomb-the-Russians-now', wheelchair-bound, strategic advisor to the US President.

4 On Oedipus' birth, his father Laius, King of Thebes, was mightily disturbed by a prophecy that his newly born son would kill him and take over his throne. Thus, he had the days' old Oedipus removed from the palace by a shepherd with clear instructions to kill the boy. However, the shepherd took mercy on the young prince and raised him as his own child. Many years later, during a chance meeting at a crossroads, Oedipus did not recognize Laius. In perhaps the earliest recorded case of road rage, he killed him in a straight duel, thus fulfilling the prophecy. Had Laius *not* believed in it, his actions would *not* have confirmed it!

5 To see the similarity between Hobbes' argument and the *N*-person prisoner's dilemma (also known as the *tragedy of the commons*), suppose there are N players and a common asset (e.g. freedom from prosecution in the prisoners' case, a river full of fish, the village green, some common resource or property, or even more abstract 'goods' such as Peace, Trust and Benevolence). Each player can grab (that is, appropriate privately) a piece X of that public good for private use (in the prisoners' case this selfish move corresponds to 'grassing' on one's accomplice). Let us normalize the value of the public good for each individual by restricting X to the range [0,1]; where $X = 0$ means that she has abstained altogether and $X = 1$ that she has grabbed the most that is possible for a single individual to grab.
The idea here is that the greedier the players (i.e. the closer the average value of X is to 1), the greater the depletion of the public good and the less is left for all, and each, to enjoy. This idea is borne out in the following simple payoff function for person i:
$P_i = 1 - 3\mu + 2X_i$, where μ is the average choice of X in the population of players.
Note that the payoffs are normalized so that, when the public good is intact, each person enjoys 1 unit of it (that is, if everyone chooses $X = 0$, each receives $P = 1$). Secondly, individual greed will boost μ and eat into each person's private payoff (e.g. if everyone were to set $X = 1$, a payoff of zero would result for all). Undoubtedly, each of our N players shares a collective interest to resist temptation (and thus choose $X_i = 0$). Nevertheless, and herein lies the 'tragedy of the commons' or the 'paradox' in the Prisoner's Dilemma, *at the very same time* player i has a pressing private reason to set $X_i = 1$! What is this reason? It is that being anti-social pays better *regardless of what others do*. And since this applies to all, the commons are wrecked (and all the prisoners grass on one another, thus landing in gaol together).

In Hobbes' context, the public good is Peace and X translates into an act of violence on unsuspecting fellows for the purposes of thieving, controlling and overpowering. The intuition is that, in a world of doves, a hawk can do really well; and if all others act like hawks, you might as well be a hawk yourself. But if each resorts to violence, a *war of all against all* results and Peace perishes. In that hideous environment, advocating dove-morality does not help. Hobbes suggested that the only alternative to living in fear is to empower some *Leviathan* with the authority and the means to keep us *all in awe*. (He was of course referring to the King or, in our days, to the State). In this sense Hobbes was the first liberal advocate of the legitimacy of the State's absolute authority over individuals; what seems like a contradiction was resolved by showing that rational subjects would *want* to consent to Leviathan's authority.

6 The strategic structure of Rousseau's argument is similar to that of Hobbes' in the previous note, albeit it differs in one crucial detail. Instead of μ being the *average* private appropriation of the public good (i.e. the average value of X chosen within the population of players), imagine that it is the *maximum* private appropriation of the public good (i.e. the maximum value of X chosen by someone in the group). Then there is no natural tension between private and collective interest (of the sort that prevails in Hobbes' analysis). Rather, in Rousseau's 'game' the outcome hinges on whether players trust one another sufficiently not to plunder the common asset; equivalently, the outcome depends on the degree of optimism within the team. For if Jack thinks that no one will choose a high value of X, then he will *want* to choose a low X himself. (Contrast this with the opposite scenario in the Hobbesian game of the previous note.)

On the other hand, if Jack suspects that Jill, another member of the team or group, might choose a high X (perhaps because she, in turn, fears that someone else might be choosing a high X), then Jack's best bet is to choose a high X too. Indeed, the optimal strategy here is to choose a value of X equal to what you think the maximum choice of X among the rest of the group will be: *to be as committed to the common good as you think the least committed person will be.*

Rousseau's point was simple: In sharp contrast to Hobbes' views, Peace, Trust and Cooperation are not doomed when a ruthless and all-powerful *Leviathan* is looking the other way. They will flourish if persons are optimistic and share a sense of belonging to a cooperative enterprise. In that case, public-spiritedness will generate self-confirming optimism. If, on the other hand, the political process gives persons reason to think as isolated selves, the end result might be self-confirming pessimism. In conclusion, whereas Hobbes was certain that nothing good could come out of unbridled freedom, Rousseau thought that things could go either way (a first encounter with indeterminacy in games!). In the 13th entry of this volume, Amartya Sen uses a Rousseau-type game (he calls it a 'Common Assurance Game') in order to warn against the pessimism inspired in the 1950s and 1960s by the increasing popularity of the Prisoner's Dilemma. Note that Rousseau's own narration of the above revolved around a team of hunters who could either join forces to catch a stag, so that all can eat well (a feat depending on the commitment to the task of each and every member, as opposed to the average commitment), or abscond and hunt separately for smaller prey (e.g. rabbits) to be eaten individually. Thus this game is also known among game theorists as the *Stag Hunt*.

7 In Puccini's *Tosca* the heroine's lover is arrested and sentenced to die. Scarpia, the police chief, promises Tosca that he will substitute blanks for the bullets in the firing squad's rifles if she agrees to submit to his advances. She agrees, but as they embrace she stabs and kills him with a hidden dagger. However, in the same way she had 'defected' from the agreement, so had Scarpia. Real bullets were fired at Tosca's lover, culminating in the tragic death of all three protagonists (predictably, the devastated Tosca leaps to her death).

8 His example involved two farmers whose crops ripened at different times. Should the one whose crop is not ready for harvest yet help the other now, in anticipation of similar assistance from him later? What if he helps him now, at great personal cost, but then his neighbour feigns some debilitating ailment when the time to reciprocate the favour arrives?

9 Smith's analysis can also be narrated in terms of the Hobbesian game in note 6. Suppose the players are sellers in some market and X is the total (monopoly) profit they could have captured had they colluded. Now, each may try to do better for herself by lowering prices a little. X would then fall, but her individual share would rise, since customers would be flocking in, attracted by the lower prices. However, when all sellers do the same, prices collapse, X falls to zero and none of them profit. The beauty of the market, according to Smith, is that it turns greed against the sellers and therefore forces them, in pursuit of high profit, to lower both prices *and* profit to the lowest possible level. To the extent that this is Smith's vision of the Good Society (i.e. one that provides the greatest amount of commodities to the largest number of people at the lowest possible prices), the profiteering motives of the vile merchants are harnessed for the greater good of humanity. After all, we are all consumers in the end, and we ought to be grateful to the invisible

hand which, behind our backs, transforms, through the providential operation of market forces, our nastier, private motives into public virtues. From game theory's perspective, the interesting aspect of Smith's argument is that the invisible hand reduces to the logic of the prisoner's dilemma outlined in note 6.

10 Taking his cue from Adam Smith, Karl Marx argued that, once profit disappears (due to Smithian competition), the only thing capitalists can do to revive profitability is invest in capital goods; that is, in labour-saving technology. Thus, productivity rises and more is produced. However, if the value of commodities is proportional to the human labour necessary in order to produce them, then automation will reduce values. As values fall, prices follow suit. Each capitalist wants to employ as few workers (per output produced) as possible and to pay them next to nothing. However, he also wishes that all the other capitalists employ lots of workers and pay them well; otherwise, who would have the money to buy his commodities? Capitalists, as a class, would be better off if they showed self-restraint in their use of labour-saving technology and low-waged labour. However, just as in the game of note 6, each capitalist has an incentive to use more machines and lower the wage *whatever the average use of labour saving-technology and the average wage*. In the end, capitalists face falling prices and shrinking demand for their commodities. This results in under-consumption (or over-production) which, in turn, leads to crises in profitability and, therefore, to recession and occasionally depression.

11 To check that this is so, consider A's intention to choose X (> 0) on the belief that someone else in this group, call her B, will select $2X$. For this to be so, A must entertain the expectation that B thinks *mistakenly* that there is someone else, say C, who will select $4X$. Thus any decision to choose a number greater than zero is predicated on beliefs centred upon the prediction that someone (B, in this case) is acting on false predictions. In conclusion, the *only* action which does *not* need to be founded on the assumption that someone along the line will hold mistaken beliefs, is the action of selecting zero. Nash therefore 'solved' games by discarding all beliefs that lead to actions which, in turn, contradict the beliefs that brought them about.

12 For a detailed exposition of Nash's equilibrium concept, see Hargreaves-Heap and Varoufakis (2004: Ch. 2). For a reader-friendly proof of Nash's theorem see Dutta (1999).

13 An asset can be exogenous (e.g. a windfall profit or an inheritance) or, indeed, one that may be due to *their* actions, past or future (e.g. the result of some partnership, the non-labour surplus of a firm, the gains from trade in the context of a bilateral monopoly or multilateral agreements, as in the case of the World Trade Organization).

14 Of course, as we saw repeatedly in previous sections, this does not mean that agents achieve maximal utility. Very often they undermine one another (recall the *Prisoner's Dilemma* and the *Race-to-Zero*) and, hence, their utility suffers because they are pursuing it so ruthlessly!

15 For example, Jill may value the last 1% of her overall share of V in inverse proportion to her overall share of V, whereas this may not be so for Jack. Or, equivalently, once she has (say) secured a certain portion of V, Jill may fear risking disagreement, by demanding even more, more than Jack does.

16 Let Jill's and Jack's utility functions be $u(x_L)$ and $v(x_K)$ respectively. Nash predicts an agreement (x^*_L, x^*_K) such that $100 - x^*_L = x^*_K$ and $u'(x^*_L)/u(x^*_L) = v'(x^*_K)/v(x^*_K)$. The last equation says that, at the agreement, the ratio of Jill's marginal utility from her share of the 'pie' to her utility from that share, will equal Jack's ratio of marginal utility from his share of the 'pie' to his utility from that share. Simple manipulation of that equation leads also to the conclusion that the proposed agreement maximizes the product of their utilities $u(x_L) \times v(x_K)$.
Proof: Since $u'(x^*_L)/u(x^*_L) = v'(100 - x^*_L)/v(100 - x^*_L) \rightarrow u'(x_L) \times v(100 - x^*_L) = u(x^*_L) \times v'(100 - x^*_L)$. But this last equality is the first-order condition for the maximization of the product of utilities $u(x_L) \times v(x_K)$. QED.

17 Note that the following is not Nash's own proof. My proof is based on a narrative which, while analytically equivalent to Nash's, brings out the behavioural aspects of Nash's bargaining solution (for its full version see Hargreaves-Heap and Varoufakis 2004: Ch. 4).

18 For further elaboration of the link between the social contract tradition and Nash's bargaining solution, see Hargeaves-Heap and Varoufakis (2004: Chs 4–6) as well as Varoufakis (1991: Ch.7).

19 Thomas Bayes proposed a simple rule for utilizing new information in order to update expectations. Suppose we are uncertain that some event A will occur (e.g. rain today). Let our subjective probability that it will occur be p. Suppose also that we observe another event, say B, which we know to be causally related to A (e.g. a cloudy sky). How should we update p now that we have observed B? Bayes argued that we need the following information to do so: our initial estimate p (suppose it equals 1/3 that is, we expected rain with probability one-third prior

to looking at the sky); an estimate of the probability of rain if it is cloudy (let it equal 2/3); and an estimate of the probability of cloud when there is no rain (let it be 1/3). Assuming further that the probability of rain without cloud is zero, Bayes showed that, upon observing an overcast sky, our new p, call it

$$p' = \Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B|A)\Pr(A) + \Pr(B|{\sim}A)\Pr({\sim}A)}$$

20 Let us take for illustration purposes the *7-coin distribution game* in section 2.2 above. Suppose that a population of N automata play this game, two at a time. In each round, some fair coin is tossed to decide who visits the table first. Once a game is over, the players are rematched against a fresh opponent and play the same game again ad infinitum. In this evolutionary setting, we assume that players do not understand the game's structure but only mimic the behaviour that amasses more coins. Also, we assume that, at every point in time, some player will play unpredictably (the mutation mechanism). Suppose that at the outset there are two types of player: *Type 1* collects 2 coins every time she visits the table while *Type 2* collects 1 coin. In a population of *Type 2* players the player that is selected to play first will end up with 4 coins and the other one with 3. Although the population prospers (in the sense that the result is efficient each time, as no coins are wasted) and equality rules (as they collect 4 or 3 coins with the same probability), consider what happens when a player accidentally (as a result of the mutation mechanism) stumbles on a new strategy: when playing second, she collects 2 coins in the penultimate round (when there are 2 coins left on the table). Clearly, her payoffs will be greater than the remaining *Type 2* player's. Thus, the adaptation mechanism will reinforce this strategy as the remaining players copy it. Eventually, no game will proceed beyond the 5th round. By the same argument, a mutation that instructs players who choose first to take 2 coins when there are 3 coins left (that is, in the 4th round) will also prevail. And so on, until all *Type 2* players evolve into ... *Type 1* players. Will the opposite be true? What would happen if, initially, all players were of *Type 1* and a mutation instructed one of them to play *as if* she were a *Type 2* player (that is, to take 1 coin only) in one of the rounds? Would this mutant behaviour catch on? It would not, as the mutant player's payoffs would be lower than everyone else's and, for this reason, no one would copy it. In conclusion, evolutionary pressure will lead all *Type 2* players to extinction and, therefore, in an evolutionary equilibrium, everyone will behave *as if* they had read the preceding Nashian analysis (even though, in reality, they are irrational automata who mechanistically copy relatively successful behaviours).

21 In the preceding pages, in an attempt to put on display the theory at its best, the chosen games all featured a unique equilibrium. But, even there, if the games are played over and over again by the same players (indefinitely), it is possible to show that the number of equilibria tends to infinity. Game theorists refer to this result as the *Folk Theorem*.

22 See, for instance, Hollis (1996), especially the chapters 'A rational agent's gotta do what a rational agent's gotta do' and 'Honour among thieves'.

23 Note that, under (a), Jill perceives Jack as someone who is simply trying not to be unfair to himself. However, under (b), Jill thinks of Jack as someone who attempts to take advantage of her efforts to cooperate with him for the joint good.

24 Note that this remarkable transformation has been achieved without departing from instrumental rationality or introducing utilitarian altruism into the analysis.

25 See Theorem 5 above, Hargreaves-Heap and Varoufakis (2004: Ch.6) and Varoufakis (2008).

26 For a poignant newspaper story, see Hal Varian's article in the *New York Times*, 29 August 2002.

27 For examples of some interesting results, the reader may sample Camerer (1997); Camerer and Thaler (1995); Camerer (2003); Hargreaves-Heap and Varoufakis (2004).

References

Aristotle
1935 'Economics' in *Aristotle's Collected Works Vol.18*, trans. G. C. Armstrong. Cambridge, MA: Harvard University Press and London: William Heinemann.

Aumann, R., and S. Hart (eds)
1992 *Handbook of game theory*. Amsterdam: North-Holland.

Camerer, C.
1997 'Progress in behavioral game theory'. *Journal of Economic Perspectives* 11: 167–188.

Camerer, C.
2003 *Behavioral game theory: Experiments on strategic interaction*. Princeton: Princeton University Press.

Camerer, C., and H. Thaler
1995 'Anomalies: Ultimatum, dictators and manners'. *Journal of Economic Perspectives* 9: 209–219.

Dutta, P.
1999 *Strategies and games: Theory and practice*. Place?: MIT Press.

Elster, J.
1982 'Marxism, functionalism and game theory'. *Theory and Society* 11: 453–482.

Greene, B.
2000 *The elegant universe: Superstrings, hidden dimensions and the quest for the ultimate theory*. Place?: Vintage.

Hargreaves-Heap, S. and Y. Varoufakis
2004 *Game theory: A critical text*. London and New York: Routledge.

Hobbes, T.
1651/1991 *Leviathan*. R. Tuck, ed. Place?: Cambridge University Press.

Hollis, M.
1996 *Reason in action*. Cambridge: Cambridge University Press.

Hume, D.
1740/1888 *Treatise of human nature*. L. A. Selby-Bigge, ed. Oxford: Oxford University Press.

Macchiavelli, N.
1985 T*he Prince*, trans. H. Mansfield. Chicago: Chicago University Press.

Marx, K.
1967 *Capital, Vol. 1*. New York: International Publishers.

Meikle, S.
1995 *Aristotle's economic thought*. New York: Oxford University Press.

Myerson, R.
1999 'Nash equilibrium and the history of economic thought'. *Journal of Economic Literature* 37: 1067–1082.

Nash, J.
1950 The bargaining problem', *Econometrica* 18: 155-162.

Nash, J.
1953 'Two person cooperative games'. *Econometrica* 21: 128–140.

Rousseau, J.-J.
1762/1973 *The social contract*. G. Cole, ed. (together with the *Discourses*). London: Dent.

Smith, A.
1776/1976 *An inquiry into the nature and causes of the wealth of nations*. Oxford: Clarendon Press.

Varian, H.
2002 'Economic scene; Avoiding the pitfalls when economics shifts from science to engineering'. *New York Times*, 29 August 2002.

Varoufakis, Y.
1991 *Rational conflict*. Oxford: Blackwell.

Varoufakis, Y.
2008 'Capitalism according to evolutionary game theory: The impossibility of a sufficiently evolutionary model of historical change'. *Science and Society* 72: 63–94.

Yanis Varoufakis

Yanis Varoufakis (PhD) has taught Economics at the University of Athens since 2002, where he also directs UADPhilEcon (an international doctoral programme in economics). He has also taught at the University of Essex, University of East Anglia, University of Cambridge and University of Sydney, with spells at the University of Glasgow and Université Catholique de Louvain. He has published in journals such as *The Economic Journal*, *Journal of Economic Methodology*, *Erkenntnis* and *Science and Society*. His books include *Rational Conflict* (Blackwell, 1991), *Foundations of Economics* (Routledge, 1998) and *Game Theory: A critical text* (Routledge, 2004).

Address: Department of Economics, University of Athens, 14 Evripidou Street, Athens 10559, Greece.

Email: yanisv@econ.uoa.gr