YANIS VAROUFAKIS

MODERN AND POSTMODERN CHALLENGES TO GAME THEORY*

ABSTRACT. Equilibrium game theory borrows from neoclassical economics its rationality concept which it immediately puts to work in order to produce the basic results it needs for building an elaborate narrative of social interaction. This paper focuses on some recent objections to game theory's use of rationality assumptions in general, and of backward induction and subgame perfection in particular, and interprets them in the light of the postmodern critique of the grand meta-narratives which social theorists often rely on for social explanation. The paper presents a defence of game theory which seeks to accommodate the postmodern critique. However, it goes on to show that such a defence is illegitimate and claims that the problem lies with the faulty conceptualisation of the main concept on which game theory rests: that of Reason. Having established the nature of the problem, it considers three alternative interpretations (Humean, postmodern and Hegelian) of why the problem resists logical solutions and of its significance for social theory.

INTRODUCTION

The battlelines in social theory have frequently been drawn along two familiar views of human agency. First, there is the perception of the sovereign agent whose autonomous desires forge the social structures that will fulfil them. Society is, therefore, seen as the means by which the agents' ends will be instrumentally realised. Second, there is the view of an individual whose desires are the product of social structure. Even if the agent rationally pursues her objectives, she is still a plaything of social forces which she cannot control.

In this paper the above controversy is bypassed in favour of a deeper controversy between the dominant variant of modernity and its foes. To accomplish this, I focus on a well-known game in which agents have given payoffs and a unique equilibrium strategy. That a fierce controversy is engendered in a simple framework is evidence that one does not need complex social interactions in order to end up with complex social phenomena. That this controversy also has the potential of inciting clashes between Humeans, postmodernists and Hegelians, is an indication that game theory ought to be more than a search for clever strategies. Indeed, as I will argue, it should be primarily concerned with the meaning of rationality in social settings.

Erkenntnis **38**: 371–404, 1993.

The following analysis is based on one particular game (often referred to as the centipede game) although it is not too difficult to show that the main problem is pervasive in game theory. In a recent paper, Sugden (1991) demonstrates the generality of similar concerns. This paper re-evaluates an increasingly popular critique of game theory's method by constructing a sophisticated defence of game theory only to show that it is ineffective. The ensuing discussion sharpens the critique and it allows us to draw parallels between the debate on game theory and some crucial philosophical controversies. In this vein, the analysis is aimed at new interpretation rather than at new solution concepts. Part of the offered interpretation is directed at game theory itself. To give a flavour of what follows, I will propose that we interpret conventional game theory as an extremist faction of the modernity project. Defining modernity as the optimism generated by the Enlightenment and concerning the ability of Reason objectively to answer complex questions concerning nature and society, I will conclude that there are three alternatives: (a) to remain within modernity while renouncing its more extreme (equilibrium theoretical) branch, (b) to reject modernity's concepts altogether (the postmodern suggestion), and (c) to turn away from the dominant aspect of modernity towards a hitherto neglected version of it.

1. INDUCING EQUILIBRIUM BELIEFS

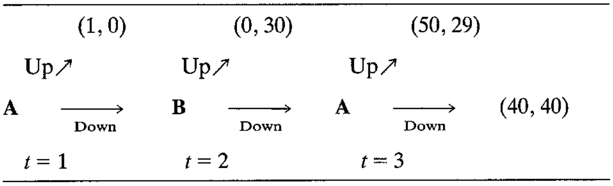
Suppose we have two individuals whom we pit against each other. We promise them a large sum of money and ask them to find some way of splitting it between them. However, we shall let them collect their reward only if they strike a deal. Furthermore, as the seconds tick away without an agreement, we continually reduce the sum in order to give them an incentive to agree quickly. Can we have a theory of what will happen? Game theory produces a narrative of what will happen starting with the simplest of cases. Suppose, we are urged, that the two are identical and that they know it. Not only do they share the same objectives, but, also, they are transparent so that each knows exactly what the other desires. If this is the case, game theorists assume that there is a unique outcome provided each agent is entirely rational,[1] knows that the other is entirely rational, knows that the other knows this etc. (from now on this assumption of *common knowledge rationality* will be referred to as CKR): they will instantly settle for a fifty-fifty split.[2] Once this result is obtained, game theory relaxes its assumptions progressively and tackles more complex versions of the same problem. First, it allows for differences in attitudes toward risk and rewards the (relative) risk-taker with a greater payoff, and, second, it introduces asymmetric information in order to show that the possession of more information is advantageous (see Harsanyi 1973).

Why is the above example representative of contemporary modernity? Two individuals facing each other in an instance of pure antagonism develop trains of thought which swiftly terminate the cacophony that would have arisen in a pre-modern narration of their situation. Instead of the drama of equally intelligent belligerents duelling to the last for personal gain, Reason is called upon to furnish harmony and efficacy. The fact that their rationality is common knowledge is presented as the bedrock of a uniquely rational train of thought that each will, if rational, latch on to. The clear separation of Reason from Unreason allows the theorist to view agents as identical computers running the same software with identical initial information (input) guaranteed by the assumption of perfect information. It is not surprising that they will inevitably come to the same conclusion (output) and thus agree without delay.

An immediate postmodern concern is that the computer metaphor inhibits understanding of human responses since our reasoning can be neither unique nor transparent. A more serious postmodern objection is that to talk of Reason is to talk of a term that has no concrete equivalent in social reality. It is not that agents are irrational, but that it is unclear what it means to be rational in social interactions. If this is so, the analogy with the numerical algorithm is misleading.[3] Whereas game theory treats simple social interactions with given objectives in a way that the outcome can be assumed and used later for explaining more complex situations (the *analytic-synthetic* road to explanation), postmodernity claims that the only chance of defensible choices, even in simple situations, materialises when we recognise the impossibility of understanding our reasoning by means of metaphors that devalue and oversimplify.

In a manner reminiscent of Parmenides' definition of nothingness (i.e., a radical absence of reality[4]) game theory identifies Reason with the residual left behind once Unreason has been expelled. By contrast, postmodernity claims that human thoughts are irreducible to a field where Reason is dominant. Critics of the modernity that lies behind

Table 1.



game theory have frequently accused it of having a social-less theory of individual agency, of procuring a process without a subject. This is not the postmodern position. The latter denies both the possibility of subjectivity *and* of analytically breaking down complex social interactions into simple ones before synthesising the resulting insights into a general social theory.[5] It is interesting to explore the connection between the postmodern critique which has been developed by writers on literary criticism and philosophy (and who have probably never considered game theory) and recent criticisms developed by game theorists.

Consider the interaction between A and B in Table 1.[6] Potentially, there are three stages in this game which begins with A having a choice of putting an end to it (by playing UP) or passing on the baton to B (by playing DOWN). If A chooses the latter, it is up to B to choose whether the game will proceed to $t = 3$. If it does, then A has the final say. Glancing at the payoffs, two things become clear. Both players are better off if $t = 3$ is reached than if A terminates the game at $t = 1$. On the other hand, A can see that if the game is to end at $t = 2$, rather than at $t = 3$, she would be better off putting an end to it right at the outset. Supposing that they have no way of communicating with each other either prior to the play of the game, or during it, other than through their UP/DOWN choices, is there a way of predicting with certainty what they will do?

Before answering, game theory introduces its axiom of *common knowledge rationality* CKR: A knows that B is rational, B knows that A is rational, A knows that B knows that A knows . . . that B knows . . . that A is rational – *ad infinitum*. And what does 'rational' mean? It means that if there is a strategy which maximises one's payoffs, one will recognise it and adopt it. So, A tries to work out whether it is better to play DOWN during $t = 1$, thus giving B the option of ending or continuing the game, or to play UP collect payoff 1 and end it there

Table 2. The logic of the equilibrium strategy (A → UP at $t = 1$)

Assumptions:
(a) AbB is rational
(b) AbBbA is rational

Fundamental conjectures:
(c) $t = 3$ *C1:* A → UP
(d) $t = 2$ *C2:* Bb*C1* thus B → UP
(e) $t = 1$ *C3:* Ab*C2* thus A → UP

A's composite conjecture inducing the equilibrium outcome:

(f) AbBbA → UP at $t = 3 \Rightarrow$ A → UP at $t = 1$

where *b* and → denote the verbs *believes* and *plays* respectively

and then. Her decision hinges on what she expects B to do at $t = 2$. If she is convinced that B will choose UP, then she ought to give him no opportunity of doing so since her payoff would be 0 compared to the 1 from playing UP at $t = 1$. If, on the other hand, she expects him to play DOWN, then she should let him do this because reaching $t = 3$ will endow her with payoff 50. However, game theory claims that this is an expectation she will never entertain.

Player A attempts to predict B's thoughts at $t = 2$ by considering what she would have done had she been in his shoes. A thinks:

> B will play UP at $t = 2$ if he expects that his payoff from doing so, i.e. 30, is greater than what he can rationally anticipate at $t = 3$. At $t = 3$ I am the one who does the choosing and I will clearly play UP leaving B with payoff 29. Since this is less than what he will get from ending the game at $t = 2$, it is silly of me to expect him to play anything other than UP. Thus, the conclusion that $t = 3$ will not be reached leads me to the conviction that I am better off by playing UP at $t = 1$.

The above logic is based on backward induction and generates the unique equilibrium set of beliefs that allows A to come to a conclusion about the best course of action.[7] As Table 2 shows, it unfolds backwards, beginning with a conjecture at $t = 3$ which leads to A's final conjecture at $t = 1$. The process that takes A from (c) to (e) is underpinned by CKR, i.e. (a) and (b).

Earlier I referred to the game theoretic predilection to the assumption that two agents with identical payoffs, rationality and information are, ontologically, identical. They are to be seen as 'running' on identical algorithms or software and coming to the same conclusion. If this is correct, then A can replicate B's thoughts perfectly since she can put

Table 3. The strategy of rational agents according to Table 2

STEP 1	Compute P_3 as your maximum payoff at $t = 3$ in the following manner: if you are player A, choose P_3 as the largest payoff; if you are player B, choose P_3 as the payoff you will collect when player A chooses her largest payoff
STEP 2	Compute P_2 as your payoff at $t = 2$ if the game is ended there
STEP 3	If you are player A go to STEP 6; otherwise continue
STEP 4	If $P_2 < P_3$ play DOWN at $t = 2$; if $P_2 > P_3$ play UP at $t = 2$
STEP 5	STOP
STEP 6	Compute P_1 as your payoff at $t = 1$ if the game is ended there
STEP 7	Play DOWN at $t = 1$ if either (a) at STEP 4 the decision is to play DOWN *and* $P_1 < P_3$, or (b) at STEP 4 the decision is to play UP *and* $P_1 < P_2$. Otherwise play UP

herself in his shoes by pretending that his payoffs are her own. This is what allows game theorists to assume that the division game described earlier is trivial and also that the passage from (c) to (e) and (f) in Table 2 ought to be automatically accepted.

To sum up, Table 2 is a good example of how the dominant modernity lurking behind game theory analyses a simple interaction between two agents, of how it breaks beliefs down to their elemental components and uses induction in order to put back together a string of conjectures leading to an equilibrium result.

2. THE CHALLENGE

Game theory establishes a rationalist vision of order which promises to 'solve' complex social interactions. Effectively, it turns social phenomena into the subject of natural scientific discourse. The logic of backward induction in Table 2 is a simple example of this. It begins with assumptions concerning the rationality of agents and derives their unique thought process. Table 3 converts this logic into a computer program which rational agents must follow, at least according to mainstream game theory.

The point to note here is that CKR renders the above program common property. It is presented as the uniquely rational sequence of conjectures that one must have when one seeks to maximise one's payoffs. Agents are assumed to recognise in it the optimal algorithm before they are allocated the role of A or B. Thus, they have worked out its logic in advance and expect that a rational A will play UP during the first stage if the payoffs are as in Table 1. Not surprisingly, when roles are finally assigned, whoever gets the part of A plays UP instantly. CKR does for equilibrium game theory's view of Reason what the veil of ignorance does for John Rawls' concept of justice: it defines it via a process of de-personalisation. The second point to note is that the adoption of this program requires that Reason is a means by which agents (as well as game theorists) convert an expectation into a conviction. For example, at stage 1 player A is facing a choice between a certain reward (payoff 1 if she plays UP) and a conjecture concerning what she will end up with if she plays DOWN. Backward induction, faithfully reproduced in Tables 2 and 3, turns this conjecture into the conviction that, were she to play DOWN, her payoff would be 0.

It is now time to explain why the above is highly problematic. The critique of backward induction which follows has been around for some time[8] but its impact has not been felt outside the narrow circles that produced it. Nevertheless, it is an important critique with repercussions for the way social theorists incorporate game theory in their models but also because it allows us to place the debates between game theorists within the larger debates in social theory. It begins with a devious thought that may cross A's mind:

> I understand Table 2 well and I agree with assumptions (a) and (b). Therefore, I see why its logic should lead me to the conclusion that UP at $t = 1$ is the only sensible strategy for me. However, what if I choose to defy it?

For A rationally to pursue this thought, she must be able to support it by a consistent train of conjectures similar in structure to those in Table 2. Table 4 presents such a sequence. The question is whether it is rational to entertain such conjectures.

Game theory's conventional response is that such thoughts are incompatible with rationality. The deviant logic in Table 4 is axiomatically ruled out on the basis of CKR. If agents take the Table 3 algorithm to be the best way of playing the game, then subjective probabilities p and q (see Table 4) must be zero at all points in logical time. If this is so, a rational A who is linked mentally via CKR to a rational B will never contemplate any strategy other than UP at $t = 1$. But why should players believe that the Table 3 algorithm is the one they ought to follow?

If the CKR is a necessary condition for dominance of Table 3, but is a condition that rationality itself cannot support, then there may be an opening for Table 4. Let us define a deviant choice as one which

Table 4. The logic of the deviant strategy (A → DOWN at $t = 1$)

Assumptions:
(a) AbB is rational with probability $1 - w = 1$
BbA is rational with probability $1 - p = 1$
(b) A and B know (a)
(c) If at $t = 2$ p were to equal 1 then: BbA → DOWN at $t = 3$

Definitions:
(d) Let $p' > 0$ be the probability belief of B at time $t = 2$ that would induce B to play DOWN at $t = 2$
(e) Let q be A's probability belief that p exceeds p', i.e. $q = \Pr(p > p')$
(f) Let $q' > 0$ be the probability belief of A at time $t = 1$ that would induce A to play DOWN at $t = 1$

Fundamental conjectures:
(g) A believes that if she defies the logic of backward induction and plays DOWN at $t = 1$, then B will revise p upwards at $t = 2$
(h) $q > q'$ at $t = 1$ and, therefore, A → DOWN

where b and → denote the verbs *believes* and *plays* respectively

goes against the equilibrium prescription of game theory but which may or may not be irrational. For example, in our game (see Table 1), if A ever played DOWN at $t = 3$, we conclude that A is (instrumentally) irrational. However, if A plays DOWN at $t = 1$, then she may or may not be irrational depending on her ability to justify her choice in terms of her objectives and beliefs. If she can justify her belief in the superiority of playing DOWN as a strategy for reaching the (50, 29) outcome, then her strategy is deviant albeit not irrational.

The logic of Table 4 is this: At $t = 1$ player A contemplates playing DOWN instead of her equilibrium strategy UP for a simple reason: she is hoping that by so doing $t = 3$ will be reached. Why? She thinks to herself:

> If B is convinced that at $t = 3$ I will play UP then he will always play UP at $t = 2$ and then we will never reach $t = 3$. Thus, if I believe that this is what he thinks, then I should choose my equilibrium strategy and play UP at $t = 1$. Indeed, according to (a) and (b) in Table 4, I know that he believes most strongly that I am rational and, therefore, he currently expects with probability 1 that, in the hypothetical case that we reach $t = 3$, I will play UP. So, at first glance I should conform with the equilibrium logic of Table 2. However, according to (b) in Table 4, this is exactly what he expects me to do. What if I do not oblige and play DOWN at $t = 1$? Surely, he must sit back and take notice.

This last thought is the gateway to the deviant logic. In trying to anticipate what B will think, agents are forced to stop operating like automata, to ditch the program in Table 3, and to start *thinking* as opposed to following formulae.[9] A continues her reflection.

> Since my choice deviates from that of the Table 2 recommendations, he will be forced to find an explanation. There are two possibilities. One is that he will think that I am irrational for not doing as Table 2 prescribes. If this is so, he will change his game plan and play DOWN at $t = 2$ expecting my irrationality to overcome my senses so that at $t = 3$ I will choose DOWN. Of course, there is the other possibility that I must reckon with. Player B may realise that this is exactly what I am thinking and refuse to believe that I am irrational simply because I have chosen irrationally. Nevertheless, all I need in order to consider playing DOWN is that B assigns a relatively low probability that I am irrational; not that he is convinced of my irrationality. Let p be the non-zero probability that he assigns to this prospect after observing my deviant choice at $t = 2$. If $p > 1/11$ (in terms of part (d) in Table 4, $p' = 1/11$), then his expected return at $t = 2$ from playing DOWN exceeds that from UP, therefore giving him a strong incentive to deviate from his equilibrium strategy too, i.e. play DOWN at $t = 2$. So, I conclude that if my defiance of the logic of Table 2 makes him think with probability 1/11 that I am irrational then it may, after all, make sense for me to play DOWN at $t = 1$ since there is now a realistic chance of getting 50 at $t = 3$ rather than 1 at $t = 1$.

We have come full circle. Player A accepts the assumption that B believes her to be rational with probability 1 but is prepared creatively to explore the thought that deviant behaviour must make those who *ex ante* rule out the possibility that their opponent is irrational to suspect *ex post* that this may not be so. She concludes that following her explicitly deviant behaviour, if B's *ex post* belief in her irrationality becomes positive (1/11 in our example), it may make sense to behave in a way that game theorists would consider irrational. More precisely, if A expects p to exceed 1/11 with probability a touch over 1/50 – i.e. if A expects that there is a 1/50 probability that her deviance at $t = 1$ will make B think that she is irrational with probability at least 1/11 – then her expected returns from playing DOWN in defiance of game theory's logic are greatest. Hence, part (g) in Table 4.[10]

Tables 2 and 4 offer alternative logics that A can choose from. Can they be equally valid? Game theorists favour the equilibrium story on the basis that it is uniquely compatible with CKR.[11] Under this type of common knowledge, player B will never update p upwards if A chooses her deviant strategy at $t = 1$ and, therefore, player A will never entertain a subjective probability q that exceeds 0. But this is too strong. As Pettit and Sugden (1989) have shown, a subtle difference in how we interpret shared rationality can change all this. All we need is to treat shared rationality as something agents believe in rather than as an immutable axiom. If we assume that agents *believe* that irrationality is absent at all orders of belief, instead of axiomatically dismissing any possibility of doubt concerning the presence of irrationality, then the deviant strategy is given a chance. The difference becomes apparent when we look at part (g) in Table 4 and compare it with parts (c), (d) and (e) in Table 2. In the former case a deviation from what is deemed to be rational behaviour has the potential of making B wonder whether his cast iron belief in A's rationality is well founded. In the latter case, by contrast, A and B follow the predetermined program in Table 3 since no deviation from the equilibrium scenario will make them wonder about the correctness of their conceptualisation of the game. Thus, Table 2 requires that, once rationality is assumed, players do what table 3 tells them *regardless of whether their opponents choose in the manner that Table 2 predicts*.

In summary, the point of contention seems to revolve around the agents' subjective beliefs. Game theory leans on CKR in order to rule out any uncertainty about the beliefs of one's opponent. Thus, it reduces the set of optimal strategies to the one in Table 2 and does not concern itself further with the prospect of rational deviations. If we choose a slightly ammended version of common knowledge which allows agents to re-think their conviction concerning the absence of irrationality once they observe deviations from the Table 3 algorithm, then deviance can be shown to be rational. Another way of conceiving our theoretical dilemma is this: under CKR agents are incapable of forming views about what they ought to do in the future if they find themselves at a part of the game-tree that CKR would not have allowed. They do not need to do so because CKR axiomatically assumes that no such trespassing ought to be considered. But when it is considered, agents may conclude that it is in their interest to abandon the equilibrium path. Indeed, would they not be irrational if they failed to consider all outcomes, including those that CKR deems unwise? And if the mere contemplation of these parts of the game-tree renders deviance rational (though not uniquely so), is this not conclusive proof that CKR is inappropriate?

Defenders of CKR may protest that the above argument suffers from the following defect: If A's deviance at $t = 1$ manages to raise B's estimation of her irrationality, then how does B predict an irrational A's behaviour at $t = 3$? And if B has problems at $t = 2$ in predicting

A's behaviour, how can we say that A's deviant strategy is rational at $t = 1$ when she cannot know how B will be thinking at $t = 2$? This is a good point. It proves beyond doubt that our players face risky decisions once CKR and the safety of the Table 2 logic are abandoned. This is, however, no proof of the irrationality of deviance; it is merely confirmation that neither the equilibrium nor the deviant strategies are uniquely rational.

In effect, when A contemplates the deviant strategy she is hoping that she can deceive her opponent. Is this rational? The answer must be that it is certainly not irrational. There is nothing in the structure of this game to suggest that an instrumentally rational agent ought to assume that she cannot out-manoeuvre her opponent. By the same token, it is also rational to think that she cannot do this. The problem with game theory and its CKR foundation is that it instils in agents' minds the belief that deception can never work. It is unclear what institution or psychological mechanism performs the same role in society.

3. A NEGATIVE DEFENCE OF THE EQUILIBRIUM APPROACH

Modernity inspired an extraordinary confidence about our ability scientifically to solve complex natural and social problems. The imposition of CKR by game theory may be thus interpreted as an attempt to consolidate modernity's spirit in games such as the one in Table 1. The previous section challenged this spirit by encouraging agents to ask questions such as: "What if I do not do what the theory suggests I ought to?" Of course this is not a question that automata can ask. And since game theory models agents *as if* they were automata such as the one in Table 3, then game theory fails to grasp this important dimension in rational agency.[12]

In this and the next section I will be presenting two lines of defence for equilibrium game theory. The first is a negative defence in the sense that it refuses seriously to consider the alternative (deviant) strategy advocated in Table 4. The second defence is much more sophisticated and follows in Section 4. Rather than ignoring the possibility of deviant play by player A at $t = 1$, the latter tries to explain it by means of an argument that is internal to game theoretical thinking.

In true modernist spirit, both defences rely on the belief that there exists a unique theory describing rational play in this game. Where they diverge is that the negative defence does not allow Reason to take more than one form *within* the unique theory whereas the latter does. Starting with the so-called *Harsanyi doctrine*, a negative defence would claim that if Reason is unique and unabridged and the two players are *equally* rational then (in the absence of asymmetric information) they must generate identical trains of thought.[13] Tables 2 and 3 provide the only thoughts compatible with this requirement. If we are to accept the logic of Table 4, the negative defence continues, then we accept the possibility that one of the two players may form expectations which are proved wrong by the play of the game.[14] But since we assumed that they are equally rational, how can we allow one of them to develop correct expectations while the other does not?

The above defence suggests that if we are to assume identical rationality then we must accept the equilibrium logic. Perhaps, this defence argues, it is not a good idea to make this assumption. Then, of course, it is not game theory that we must blame for producing a result we do not like but our assumptions. However, I do believe that this argument is untenable. For who is to say that if there are two identically rational agents involved in such an interaction, both of their trains of thought must be proved correct? To be prophetic is not a prerequisite for being rational. If, indeed, there is more than one rational train of thought, our players may form different sets of conjectures each being utterly rational. Quite naturally, one may end up with conjectures that are confirmed by the actual choice of strategies while the other does not.[15] This is not to say that one is more rational than the other.

Relating the above argument to the main theme in this paper, it seems as if game theory has a tendency to maintain that Reason is more powerful than it can ever be. The negative defence burdens it with the task of coordinating beliefs and choices when, on its own, it can do no such thing. The moment our players are told that (in the context of Table 1) their opponent is rational, they are supposed to know exactly what will happen because the thought that one may try to outwit the other never crosses their mind. If it could be demonstrated that equal rationality has this effect, then the defence would be successful. Unfortunately, what I refer to as the negative defence is based on the assumption that such a pernicious thought will not arise. Why not? Because game theorists believe that if two players are equally rational, then we cannot allow a situation where one of them out-manoeuvres the other. However for this to be true, it must be shown that rationality commands players who hold their opponents in high regard to abstain from efforts to outwit each other. What boring events world title chess championships would be if this were true!

Such cowardice cannot be synonymous with rationality, even if compatible with it.[16] It seems to me that the crux of the argument is that the negative defence demands that agents cannot distinguish between the following two statements: (i) my opponent is rational and thinks I am rational, and (ii) there exists only one train of thought that is rational to form in this game. I cannot see why (i) should necessitate (ii) if agents are equally rational. If it does not, the negative defence fails to meet the challenge of Table 4 and relies on a perception of Reason which is open to what Hegel wrote in the *Phenomenology*: 'It lives in dread of besmirching the radiance of its inner being through action and existence. In order to preserve the purity of its heart, it flees from contact with actuality and persists in a state of self-willed impotence'. The challenge of Table 4 is denied simply because it cannot be grasped.

4. A POSITIVE DEFENCE OF THE EQUILIBRIUM APPROACH

A positive defence of game theory ought to attempt to undermine the Table 4 logic by showing that something very similar to the latter can be constructed if we follow the method that gave rise to Table 2. In other words, a sophisticated game theorist would argue that the reason why the deviant strategy sounds plausible is because it has a perfectly good equilibrium foundation rather than because equilibrium theory is deficient. Thus, game theory attempts to assimilate Table 4 rather than to banish it. To do this it accepts the proposition that there may, after all, be more than one rational train of thought.

Before moving to the positive defence it is useful to look closer at a possible interpretation of the challenge to the original equilibrium theory. The latter urges player A to choose UP at $t = 1$ after looking at $t = 3$ and projecting the decision she would have made at that stage onto player B at $t = 2$ and then back onto herself at $t = 1$. In a sense, player A is asked to 'observe' what she would have done at t = 3, induce from that what B will do at $t = 2$ and further induce what she ought to do at $t = 1$. Whether this induction is appropriate or not depends on the projectibility of the conclusions derived from an analysis of stages 2 and 3 in *isolation from the rest of the game*, onto stage 1.

Game theory uses the CKR assumption in order to ensure that the compartmentalisation of the game into subgames separately to be examined is uniquely legitimate. However, by ignoring the projectibility of conjectures from one subgame to another it neglects an important aspect of rational induction.

Say player A is about to choose her strategy at $t = 1$. According to backward induction, she looks at $t = 3$ first and thus illuminates her current choice. Game theory identifies the ability to 'induce' in this manner a unique rationale with rationality. But is induction invariably trustworthy? For instance, she may ponder the proposition that, in logical time, all stages of the game precede $t = 1$. By induction, may she conclude that all stages of the game will share that trait? This conclusion would lead her to believe that the game will never start since $t = 1$ cannot eventuate. Taken further, a second level induction, an induction about such inductions, tells A that such inductions are always wrong. Should she now believe that the game has started a long time ago since there can never be a stage not preceded by another stage of the same game? Quite clearly, a blind application of induction does not conduce intelligent thoughts. We need something more before we resolve that a trait characterising one stage of the game is projectible onto another.[17]

Some traits command confident expectation of continuance from one stage to another and some do not. We do not expect the trait 'being prior to $t = 1$ in logical time' to carry forward to past moments without end. How do we know whether a trait is projectible or not? A phenomenon that is immediately noticeable and has recognisable form is potentially projectible by means of induction – e.g. a sunset is projectible from one day to another. Postmodern thinkers attach a great deal of importance to language. They would argue that a trait is projectible if there is a word for it that reveals, rather than hides, its true meaning. Game theory, on the other hand, derives the logic of Table 2, and pushes hard for it to be recognised as a uniquely rational logic, on the basis of an induction without establishing the projectibility of the main trait that is being carried from $t = 3$ to $t = 1$. The critique in Section 2 refuses to accept that the meaning of the word 'rationality' is clear enough to sanction unconditionally the kind of induction required for the generation of Table 2. Therefore, in circumstances of a truly interactive game, the trait 'rational choice at $t = i$' [where $i = 3, 2, 1$] is as unprojectible as the trait 'being prior to $t = 1$ in logical time' above.[18]

The result is that Table 2 cannot represent the uniquely rational train of thought. This is a familiar postmodern view regarding modernist theories which are accused of mistaking analogies for concepts.[19] In our case, game theory mistakes the consistency which results from analogous behaviour (such as that prescribed by Table 3) for rationality.

A positive defence of game theoretic orthodoxy must begin with a humble admission that Table 2 is only an embarkation point and not the destination of the equilibrium narrative. The CKR assumption should then be interpreted as an initial assumption to be relaxed soon after the theory has got off the ground. The refined game theorist must admit that there are different ways of conceptualising the game of Table 1 and that it is unwise to *assume* that a rational player A will never play DOWN at $t = 1$. However, the positive defence must insist that, if we are to understand what happens when two equally intelligent players participate in this game, we must utilise the tools of equilibrium analysis even if we respect its earlier conclusion based on quite restrictive rationality assumptions. Kreps et al (1982) offer a good basis on which to build such a defence.

The first leg of a positive defence would be to modify the CKR assumption as stated in parts (a) and (b) of Table 2. In its stead, it would place the assumption that players may now suspect their opponent to be irrational (see (a) and (b) in Table 5). Furthermore, it allows for more than one kind of rationality so that player A may be rational and yet not conceptualise the game according to Table 2; let me label the latter the Reason of Backward Induction RBI. After making these concessions, the positive defence procures its rationalisation of the deviant strategy on game theoretical grounds. Central to it is the diversity of logics which a rational player may adopt as well as the possibility of flagrant irrationality. In Table 5 player B expects player A to be irrational with probability p and is convinced that an irrational A will always choose DOWN at $t = 3$.[20] Also, if he thinks she is rational he does not immediately assume that she will adopt RBI and the Table 2 logic. He expects a rational A to deviate from RBI with probability $1 - r$ (see part (h)) due to the adoption of an alternative mode of reasoning. What kind of reasoning is that? It depends on how open-minded the theorist is. In the most stark of interpretations, to shun RBI is identified with irrationality and $1 - r$ becomes the probability of behaving irrationally with a view to confusing player B. Alternatively, game theorists may wish to allow $1 - r$ to be the probability with which player A espouses either the logic of Table 4 or some other logic without requiring that there is a hierarchy of logics with RBI at its pinnacle. Indeed, if they are keen to show that the critique in Section 2 is a special case of equilibrium logic, then they must accept that RBI is just one of many equally admissible conceptualisations. Later I will be arguing that this last claim is reminiscent of the postmodern challenge to Reason.

In the second leg of the defence we find the plausible argument that there are three reasons why player A may choose DOWN at $t = 1$ against the advice of RBI.[21] Firstly, she may play DOWN because she is irrational. Secondly, although rational, she may not subscribe to the logic presented by RBI and Table 2. Thirdly, she may be rational *and* subscribe to the RBI (and the logic of Table 2) but, nevertheless, attempt to confuse player B through her choice at $t = 1$ so that B plays DOWN at $t = 2$ giving her a chance to reap the highest payoff at $t = 3$. Of course, observation can never help B distinguish between the second and the third reasons. If A is irrational then, potentially, this will be revealed at $t = 3$ where she will choose DOWN. On the other hand, if she is rational and plays DOWN at $t = 1$, then her particular version of rationality will never be revealed via her choices as she will invariably play UP at $t = 3$. Therefore, player B lumps the second and third reasons for a rational A playing DOWN at $t = 1$ under one category and attaches to this event probability $1 - r$. In any case, this is not a serious conceptual problem since one can argue that to doubt RBI is conceptually identical to not doubting it and yet rationally to choose to evade it.

Let us now explore the interdependence between agents' beliefs and choices. Table 6 captures the possible states for player A at $t = 1$ as perceived by player B. Player B expects A to be rational with probability $1 - p$, in which case she may adopt RBI with probability r or choose an alternative logic with probability $1 - r$, or to be irrational with probability p. Probability s relates the likelihood that an irrational A will choose the deviant strategy at $t = 1$.

Suppose now that player A plays DOWN at $t = 1$. What should B think? As in Table 4, he will immediately update his probabilistic expectation that A is irrational taking into account the possibility that she may simply be using an alternative reasoning to RBI (such as the one in Table 4). Bayes' rule recommends the following consistent updating mechanism once A's behaviour is observed at $t = 1$.

$$\Pr(\text{A is irrational} \mid \text{A} \rightarrow \text{DOWN at } t = 1)\ \{\equiv p_2\} =$$

Table 5. The positive defence: a game theoretic explanation of the deviant strategy

Assumptions:
(a) AbB is rational with probability $1 - w \leqslant 1$
(b) BbA is rational with probability $1 - p \leqslant 1$
(c) A and B know (a) and (b)
(d) If at $t = 2$ p were to equal 1 then: B expects with certainty A to play DOWN at $t = 3$ (see footnotes 8 and 15)

Definitions:
(e) Let $p' > 0$ be the probability belief of B at $t = 2$ that would induce B to play DOWN at $t = 2$
(f) Let q be A's probability belief that, at $t = 2$, $p > p'$, i.e. $q = \Pr(p > p')$ if she plays DOWN at $t = 1$
(g) Let q' be the probability belief of A at time $t = 1$ that would induce A to play DOWN at $t = 1$
(h) Let $1 - r$ be the probability belief of B that A, if *rational*, will adopt the RBI reasoning (i.e. the reasoning of Table 2)
(i) Let s be the probability belief of B that A → DOWN at $t = 1$ *when* A *is irrational*

Fundamental conjectures:
(j) A rational player A believes that if she defies the logic of backward induction (RBI) and plays DOWN at $t = 1$, then B will revise p upwards at $t = 2$ using Bayes' rule (see Equation (1))
(k) $q > q'$ at $t = 1$ and, therefore, A → DOWN

where b and → denote the verbs *believes* and *plays* respectively

Table 6.

Player B's conjecture about player A at $t = 1$

			r	**UP**	$(1 - p_1)r$
	$1 - p_1$	Rational	↗		
	↗		→ $1 - r$	**DOWN**	$(1 - p_1)(1 - r)$
A					
	↘		$1 - s$	**UP**	$p_1(1 - s)$
	p_1	Irrational	↗		
			→ s	**DOWN**	$p_1 s$
				A's choice at $t = 1$	B's subj. prob. belief for each outcome at $t = 1$

$$\Pr(\text{A} \rightarrow \text{DOWN at } t = 1 \cap \text{A is irrational})/\Pr(\text{A} \rightarrow \text{DOWN at } t = 1)$$

[where the subscripts of p correspond to the time period at which these beliefs are formed].

From Table 6, it follows that the above updating mechanism can be re-written as

$$p_2 = \frac{p_1 s}{(1-p_1)(1-r)+p_1 s}. \tag{1}$$

Hence, given values for r and s, both players can work out how an initial belief that A is irrational will be updated if A plays DOWN at $t = 1$. Moreover, they both know the value of p' (i.e.. the degree of conviction at $t = 2$ that A is irrational so that B wishes to play DOWN at $t = 2$) and they can work out whether it would make sense for a rational A to play DOWN in order to ensure that p reaches p' (i.e. so that deviant behaviour at $t = 1$ pushes p_2 up to the level of p'). In our particular game in Table 1, p' equals 1/11. Given an initial probabilistic belief by A that B is irrational equal to p_1, I re-write (1) as:

$$\frac{1-p_1}{p_1} = \frac{1}{\gamma}\frac{1-p_2}{p_2} \tag{2}$$

where $\gamma = (1-r)/s$.[22]

Our players know that if 'A → DOWN at $t = 1$' is to create a significant degree of doubt in B's mind concerning A's rationality, p_2 must reach at least 1/11. Supposing that, for example, $1 - r = 1/4$ and $s = 3/4$, what is the minimum probability belief at $t = 1$ with which B expects A to be irrational? Substitution into (2) yields this level as $p_1 = 1/31$. In summary, the above model tells the following story:

> If $p_1 > 1/31$ then A knows that B will be prepared to risk playing DOWN at $t = 2$ if she plays DOWN at $t = 1$.[23]
> If $p_1 < 1/31$ then A knows that B cannot be made to feel with sufficient strength that A is irrational. So, unless A is irrational she will play UP at $t = 1$.
> If $p_1 = 1/31$ then A is indifferent between the two strategies at $t = 1$ and randomises. If the outcome of this randomisation is DOWN then p_2 will (by Equation (2)) equal 1/11 and B will also become indifferent between his options at $t = 2$. Thus, he will also randomise.

We have come to the end of an impressive defence of game theory built on standard game theoretical concepts developed by, amongst others, David Kreps and Robert Wilson, Paul Milgrom and John Roberts (see Kreps et al., 1982). It claims that once the CKR assumption is dropped (i.e. once $p_1 > 0$), player A may attempt creatively to exploit the fact that her opponent does not know if she is rational at all or, if rational, what kind of rationality she espouses. In addition, she may have an incentive to defy RBI *even though she initially conceived of the game in terms of RBI*! Turning the tables on the challenge of Section 2, game theory seeks to explain internally the logic of Table 4 by means of Table 5.

The only striking difference between this sophisticated narrative and that of Table 4 is that the latter begins with an assumption in common and absolute belief in each other's rationality, whereas the former requires that B experiences at least a little bit of uncertainty concerning A's thoughts. One could construct a claim that Table 4 is superior to the above as an explanation of deviant play at $t = 1$ because it allows deviant thoughts even when both players are convinced that they share the same reasoning. However, on its own this would be a thin claim. For the defenders of game theory could retort that, in view of the conclusions of Table 4, no rational player can be certain that she knows the reasoning that her opponent will employ. Thus, assumptions (a) and (b) in Table 4 are not the assumptions that rational players would wish to make and, consequently, it makes more sense to relax the CKR assumption instead.

A Repudiation of the Positive Defence of Equilibrium Theory

At the centre of the positive defence we find Bayes' rule. It provides the link which was missing from Table 4 and allows the conclusions of the logic in that table to hold without compromising the logic of equilibrium. Its role is to update B's initial concern about A's possible irrationality after A plays DOWN at $t = 1$. However, I wish to argue that its use in this context is so fraught with problems that we (and our players) are better off without it.[24] If this turns out to be sound advice, we will return to Table 4 and the equilibrium defence will have been fruitless as there will be no unique (equilibrium) story to tell about how our players process the information that deviance at $t = 1$ furnishes.

What conditions must hold for Bayes' rule to be operational? Assuming that A and B share the same rationality (i.e. RBI), player B must know the values of p_1, r and s. Then, B must believe that player A knows that he knows these probabilities, which requires that A and

B have exactly the same subjective probabilities on p_1, r and s. If such convergence of minds is not achieved, player B will not be able to use (1) and A will not expect him to. Let us take these subjective probabilities one at a time:

A&B have the same expectation s i.e. they have somehow homed in on the same value of the probability that an irrational person will play DOWN at $t = 1$. But is this a reasonable deduction from the assumption that A and B are both rational? Surely the point about irrational or stupid agents is that rational agents *do not* understand them. Even if one is convinced one can predict irrational behaviour (i.e. form a sound estimate of s) how can one be sure that another rational agent will form *exactly* the same estimates? And what happened to the newly found open-mindedness which would allow for more than one kind of rationalisation? If this concession is genuine, then surely there must be more than one commentary on irrationality thus giving rise to a plethora of predictions on s and wrecking Bayes' rule.

A&B share the same value of r i.e. if A and B are both rational (regardless of the particular form of rationality they subscribe to), B knows the probability with which A will play DOWN at $t = 1$. In effect, the positive defence postulates the existence of a unique theory by which player B can predict or explain the behaviour of someone whose reasoning he does not share. But how is this possible when there are many possible modes of reasoning? Moreover, even when they share the same r, how can A be absolutely certain that this is so? For that is exactly what is required before Bayes' rule can function.

A&B share the same value of p_1 i.e. a rational player A must know exactly B's subjective probability assessment that A is irrational and must know that B knows that! That they are rational when asked to form this identical belief is no guarantee that they will form it. Once more the assumption of rationality is asked to do too much.

Criticism 1. The positive equilibrium defence refurbishes the common knowledge of rationality assumption (CKR) only this time it is common knowledge of p_1, r and s – not of rationality as in Table 2. Since no one can demonstrate that equally rational agents ought to trust each other to have the same subjective beliefs p_1, r and s, it would be utterly irrational of them to act in a way that vindicates the theory proposed by the *positive defence of equilibrium analysis. Therefore, we conclude that the challenge posed by Table* 4 *has not been met by game theory.*

The standard reply by game theorists which the above criticism shall occasion is that if we want agents to entertain different expectations, then there is no real problem. We assume that this is the case, equip them with probability density functions which capture their uncertainty about each other's subjective beliefs and we derive a complex asymmetric information model that addresses the above concerns. However, this would muddy the waters unnecessarily. There is no gain to be had if a problem is elevated to a higher level of complexity without being solved. If an equilibrium model with asymmetric subjective beliefs is to work, we must assume that the probability density functions of one player are known with certainty by another. So, instead of demanding that agents use the same value of p_1, r and s, the positive defence now demands that they are certain that they have the same probability density functions over different values of p_1, r and s. But why would they feel confident that this is so? Such confidence would be unacceptable on the grounds of rationality.[26]

In order to avoid the danger of getting bogged down in a pointless argument about higher order probabilistic conjectures, and whether or not they should be in equilibrium, let me make a concession that I do not have to make and yet demonstrate that the common knowledge the positive defence depends upon is implausible. Suppose that A and B are equally rational and have at their disposal exactly the same values of p_1, r and s, as the positive defence assumes (i.e. for the moment disregard Criticism 1). Some unspecified process leads them both to the conclusion that $1 - r = 1/4$, $s = 3/4$ and $p_1 = 1/31$. The question then is: what will A and B do given these shared beliefs? Will they have an incentive to move away from them? According to the equilibrium story, if player A is rational she must randomise at $t = 1$.[27] If the outcome of this randomisation is DOWN then Equation (1) yields $p_2 = 1/11$ and player B is forced to randomise at $t = 2$ too. This is a knife-edge situation where neither has an equilibrium pure strategy and where each has to resort to an equilibrium mixed strategy – i.e. to randomising. Is this what they will do?

Player A may believe that B will stick to the above scenario. If she does, she has no overwhelming reason for playing DOWN, UP or for randomising. So, why should she randomise? Her expected returns are the same whatever she does and, hence, she may choose one of the two strategies with certainty or indeed choose to mix them in any which way she feels like. Suppose that for some unspecified reason she contemplates playing DOWN.

DEVIANT THOUGHT 1 (DT1). A decides to set $r = 0$.

Naturally, DT1 is not an equilibrium decision since only $r = 3/4$ would ensure that her choice will be in equilibrium with B's conjectures. On the other hand, it is not a foolish decision either since whatever r is set equal to, her expected returns are the same provided B believes r to equal 3/4. To put it differently, A has no incentive to stick to $r = 3/4$ even if this is the value she initially entertained. Thus, if she espouses DT1, she will choose DOWN at $t = 1$. Now, what if B thinks that there is a tiny probability that A has adopted DT1? He will immediately place $r = 0$ in (2) and derive a new probability estimate concerning A's irrationality (i.e. a value for p_2) that is below 1/11 and is incapable of motivating him to play DOWN at $t = 2$. This is captured by the second deviant thought:

DEVIANT THOUGHT 2 (DT2). B anticipates DT1 and if A → DOWN at $t = 1$, B → UP at $t = 2$.

Not surprisingly, a string of deviant thoughts may follow. Player A may anticipate that if she plays DOWN then DT2 will emerge in the mind of B and she may, therefore, set $r = 1$ since she expects B to play DOWN at $t = 2$.

DEVIANT THOUGHT 3 (DT3). A anticipates DT2 and sets $r = 1$.

If player B expects DT3 to infiltrate A's thoughts, then we move to DT4:

DEVIANT THOUGHT 4 (DT4). B anticipates DT3 and sets $p_2 = 1$ if A plays DOWN at $t = 1$. Hence, B will be prepared to play DOWN at $t = 2$.

If player A thinks that playing DOWN at $t = 1$ will give rise to DT4, then she will develop DT5:

DEVIANT THOUGHT 5 (DT5). A anticipates DT4 and sets $r = 0$.

And so on.

Criticism 2. It is not only that players will converge on the same subjective probabilities by accident alone but, moreover, that they may busily develop thoughts which will ensure the impossibility of such symmetry.

Since the actual outcome of the game will depend on which thought each player terminates his or her climb up the ladder of conjectures, we cannot predict what either of them will do. There is no optimal stopping rule when one enters deviant trains of thought and for this reason any equilibrium story (including those postulating probability expectations over the two strategies of each player) is inappropriate.[28] All we can safely say is that if A stops at DT3, then Table 2 applies. If, on the other hand, she reaches DT5, then Table 4 aptly describes her thoughts.

I predict two objections to Criticism 2. First, DT1 may be denied on the grounds that there is no reason why it is more likely to develop than, say, DT1′: A sets $r = 1$. However, in this case player B may anticipate DT1′ and move directly to DT4 setting $p_2 = 1$. The point is that at $t = 1$ anything goes whichever deviant thought arises first. The second objection is that Criticism 2 applies only when p_1, r and s are such that $p_2 = 1/11$. Although this is correct, this case is too important to dismiss as an exception that confirms a rule. Since A's choice at $t = 1$ will depend on whether she expects or not $p_2 > 1/11$, it is crucial for the positive defence that there exists *one* combination of p_1, r and s such that $p_2 = 1/11$ so that A is made indifferent between UP and DOWN at $t = 1$. Otherwise there is no clear demarcation between the case where A will rationally play the deviant strategy and the case where she will not. And without this demarcation, there can be no equilibrium defence of game theory from the challenge of Section 2. In conclusion, Criticism 2 reinforces the claim of Criticism 1 that there can be no tenable equilibrium theory of what will happen if A and B are gifted with equal amounts of Reason.[29]

At this stage it is helpful to summarise the argument in simple terms. Table 4 introduced the possibility that agents will contemplate a risky strategy. The positive defence tried to explain this as an equilibrium strategy. However, the moment such a strategy is contemplated, there is no equilibrium solution. When the Table 4 strategy is considered, an agent's choice depends on subjective judgments about another agent who must himself make subjective judgments about her earlier and future behaviour. The nature of agents' belief formation being subjective, it undermines the derivation of equilibrium solutions. Quite clearly, equilibrium theory survives only if somewhere along the line we *assume* an equilibrium outcome. Its positive defence, if it is to remain erect, needs to be underpinned with the hidden statement: 'let us assume that equilibrium theory is correct'. But this would be equivalent to the negative defence in Section 3!

5. POSTMODERN, HUMEAN AND DIALECTICAL INTERVENTIONS

An eagerness to unravel logically complex social phenomena is a commendable characteristic of modernity. The problem is that, along the way, its ambition often sweeps unresolved questions under the carpet in search of short cuts. Postmodern thinkers have questioned the concepts that modernity uses on the grounds that they are flimsy analogies rather than concepts. One of the concepts that they challenge is that of Reason. Those who are concerned about game theory's rationality postulates may find the postmodern critique useful. Looking at the preceding arguments through a postmodern prism, one interpretation of the discussion in Sections 2, 3 and 4 is that CKR is an extreme form of modernity; a byproduct of an illegitimate, yet strong, ambition to select one of A's two strategies at $t = 1$ as uniquely playable by rational agents. Postmodernists would recognise in CKR (and Tables 2, 3) the same modernist tendencies that they disparage in literary criticism, politics and philosophy (see Derrida 1978, Norris 1985, Lyotard 1984).

It is, however, perfectly admissible to accept the critique of game theory without abandoning modernity. In a Humean sense, Reason is the slave of passions (that is, the payoffs in our game) which motivate choices and acts as the disinterested judge who weighs the merits of the various options but does not pass judgment on the desires themselves (in the same way that a judge does not question the law). If desires under-determine choice Reason is not to blame for the resulting indeterminacy. As Aristotle put it in *Nicomachean Ethics*, the rules of the undetermined are themselves undetermined. Thus, the critique of game theory does not challenge modernity as such but only an extreme version of it (i.e. equilibrium game theory) which wants a determinate solution so badly that it contrives rationality concepts (CKR, RBI etc.) that are not supported by Reason. Humean instrumental Reason offers no guidance to A and B at $t = 1, 2$ because no choice is uniquely rational. What it can do is to suggest that, in such circumstances, the solution lies in convention. Conventions help agents make sense of logically indeterminate situations although no convention in itself can be understood in terms of its rationality. If we wish to understand how they are formed, we need to look at their evolutionary stability. Further, if we wish to explain why they become stable, a Humean interpretation is possible: agents develop desires that they, and others, follow the established conventions. Then, it may be rational to act in one way rather than in another as a new desire has been actuated allowing Reason to discern a uniquely rational action.

Granted that modernity is not directly challenged by Sections 2 and 4, it is worthwhile to follow the postmodern critique of it a little further. Hume shares with game theory a perception of Reason as a concept which is definable axiomatically and independently of social interaction. Postmodernity on the other hand denies the possibility that abstract signifiers such as Truth, Being, Reason signify anything concrete; that they are more than figments of our language. By contrast, we are encouraged to recognise that Reason appears as a momentary flickering of presence and absence and does not allow us a good look. The only strategy that we should contemplate is to deconstruct narratives such as the one in Table 2, to invoke Reason and then immediately to erase and fragment it. In the context of the earlier discussion, we are asked to accept the rationality of Tables 2 and 4 simultaneously not because (as the Humean would argue) Reason cannot deliberate in this case but because there is no such *thing* as Reason.

For a brief moment, the positive defence of game theory in Section 4 seems compatible with postmodernity. As it begins with a recognition that there is no unique reasoning and thus no hierarchy of logical trains of thought, one may think that postmodernity has found a mathematical expression. However, the deconstruction of that defence (see the two main criticisms in that section) reveals the inherent incompatibility between the two. For if postmodernity accepts this model, it will be using the concepts it is critical of in order to castigate them and would become vulnerable to a critique reminiscent of Heidegger's attack on Nietzsche.[30]

On a positive note, postmodernity offers an interesting answer to a question we have neglected so far. Returning to the first stage of our game, why should we discuss the rationality of various types of reasoning in terms of the backward induction logic? Why should, in view of our conclusions, label Table 4 'deviant' thus crediting Table 2 with a priority it should not have on the basis of Reason? Postmodernity has this to offer: As the Enlightenment sought scientific explanations by which to escape dogmatic certainties, natural science took it upon itself to furnish them. In the realm of natural science, the various possibilities that required analysis were states of nature and could be treated as such quite legitimately. When social phenomena were tackled, it was natural to try to apply the same logic. The problem is that human choices cannot (and should not) be treated as states of nature.[31] Take for instance backward induction. If A was to play the game not against a human agent but against an automaton whose software was describable by Table 3, then backward induction would correctly inform her that she should play UP at $t = 1$. However, when she plays against a human B, backward induction breaks down. Nevertheless, the cognitive priority that we seem to lend the backward induction logic is, according to postmodernity, an historical accident. It is simply the product of what it mockingly refers to as the 'Enlightenment episode'.

What picture of the agent is postmodernity drawing? It looks at our game and observes that at $t = 1, 2$ modernity offers no useful commentary. Only when the game reaches (if it does) $t = 3$ does modernity have an answer: A will play UP. But what kind of human subjectivity does this imply? Human creativity is responsible for creating and simultaneously undoing the backward induction logic[32] and is capable of frustrating all attempts to treat agents as automata. If we want a metaphor for understanding postmodernity's view of subjectivity, imagine the individual as a multifaceted and disintegrating interplay between selves; a series of different masks. Instrumental rationality is an empty concept if one espouses this model of men and women.

Lest we wrongly conclude that postmodernity be the only alternative to the Humean perspective, it is valuable to look at the contribution of Hegel. At the risk of oversimplification, a Hegelian interpretation of what is happening at $t = 1$ in our game is best portrayed in juxtaposition to the Humean and the postmodern views. The former evokes the image of a static Reason which, due to the inability of desires to provide it with enough information, stays on the sidelines and refuses to engage until $t = 3$ is reached, whereas the latter agrees that Reason is absent at $t = 1$ but claims that this is due to its non-existence. By contrast, Hegel would argue that Reason jumps into the fray at $t = 1$ and generates *contradictory* thought processes like those in Tables 2 and 4. And this is the rub. For it is these inconsistencies that give Reason the opportunity to enrich itself with elements of fundamentally opposed reasonings. Hegel views Reason as an evolving concept that affects the agents' experience and, in contradistinction to Hume, is affected by it.

Looking at our little game again, Hegel's dialectics suggest that at $t = 1$ Reason generates two contradictory logics (Tables 2 and 4) which are equally powerful; they are the thesis and the antithesis. The outcome is only describable in historical (as opposed to logical) time because of the logical equivalence of the two types of reasoning. However, once the game is played (and here Hegel would agree with Humeans and postmodernists that there is no way of predicting what will happen if all the information we have is in Table 1), the Reason of agents, as well as of theorists, emerges superior to what it was before they encountered this game. As it absorbs both logics (Tables 2 and 4), Reason endows us with an understanding of the game that is indescribable by one of the two tables although it is comprehensible by a synthesis of the two. Put differently, our rationality was of a lower order of development before we stumbled on this game. Generally, the more complex the social phenomena to which men and women are exposed the more advanced their Reason. Reason develops as rational agents struggle to come to grips with the maze of conjectures that social interaction (of which our game is a simple example) creates. To use Hegelian language, rationality is not to be defined axiomatically but is to be understood as a *process*.[33] If we wish to follow modernity in picturing Reason as a totality, we may still do so. However, it is not a static totality but one whose aspects are in contradiction with each other. And through this internal feud, the aspects of the totality (e.g. the conjectures in Table 2 or 4) transform not only the totality but also each other.[34]

It is quite obvious that Hegel and Hume are on modernity's side. Excepting their disparate language, what is the significant difference between the two interpretations? Both would accept the indeterminacy at $t = 1$ of our game and neither would deny Tables 2 and 4 their respective worth as equally plausible conceptualisations. I think the main difference lies in what we may describe as the byproduct of the indeterminacy. Following Sugden's (1989b, 1991) reading of Hume, the byproduct is the convention that will help agents choose in the absence of abstract logical guarantees. Reason does not shape these conventions itself although they are compatible with it. What gives rise to an impetus for their generation is the need to serve existing desires. Furthermore, in order to entrench the fledgling conventions, a new desire to abide by the evolutionary stable convention evolves – the birth of morality. It is this new desire that unlocks the problem and breaks the indeterminacy. However, the driving force behind such evolution is (a) given desires and (b) an unchanging Reason. More importantly, in this model it is impossible to pass judgment on the rationality of social conventions since Reason has had nothing to do with the selection of the particular convention. It is also futile to imagine that there is some overarching social goal that guides the evolution of conventions.

In the Hegelian perspective, however, indeterminacy bears, in addition to new desires, a new mode of reasoning – a fresh conceptualisation of one's self as one encounters the other in a social setting. Whereas in Hume indeterminacy actuates conventions and possibly new desires, in Hegel it also actuates an upgraded version of Reason. Desires, beliefs and Reason change at once when agents meet each other in a society that brings them face to face with profound contradictions. Desires and Reason are thus endogenously produced social products. The major implication is that Hegel, as opposed to Hume, sanctions judgements of the rationality of social norms that 'solve' social games on the basis of an historical analysis. When we look at past conventions we are at liberty to castigate them even if it is possible to show that agents who abided by these conventions did so because their Reason could not determine otherwise what they ought to do. Since Reason progresses in historical time, conventions that were spawned by a previous set of social circumstances, and which were perhaps compatible with agents' rationality *at that time*, may not pass the test of Reason today. While Hume's philosophy does not allow us to pass moral or political judgment on social conventions, Hegel's does.

CONCLUSION

One expects the great controversies in social theory to require a complex narrative in which to unfold. In this paper, I have focussed on a single simple game that, at first, seemed incapable of generating controversy that would go beyond the boundaries of game theory. However, it soon became obvious that, in trying to understand its structure, we were drawn to a re-assessment of the concept of rationality. The one solid conclusion of Sections 2, 3 and 4 was that there can be no unique prescription about how the game ought to be played by rational agents. Consequently, game theory's equilibrium concepts (e.g. the subgame perfect Nash equilibrium) are decidedly misleading and no amount of theoretical ingenuity can put this right (recall Section 4). Once the quest for a unique solution was abandoned, the focus shifted to a philosophical interpretation of the indeterminacy.

First came the postmodern critique of game theory. It celebrated the loss of the equilibrium solution which it sees as part of a grand narrative that constrains our understanding of human agency. Human creativity wrecks attempts to build large stories of rational choice (such as game theory) and cannot be resolved by wild-goose chases of 'concepts' such as Reason. In effect, we were advised to treat each social interaction (or game) separately and to tell local stories without trying to formulate a theory that would be applicable across games and social settings. The second interpretation was Humean. Reason is neither to blame nor to commend for the actions of our players in this game since desires underdetermine the rational choice. It is then that human creativity comes in and forges conventions which allow agents to act in the face of indeterminacy without resorting to abstract randomisations. The third interpretation is dialectical and has its roots in Hegel's philosophy. Here, creativity is also a response to indeterminacy only it creates new concepts and activates a new perception of the self as part of the process which is Reason.

The choice of interpretation has effects ranging from our conception of the meaning of rationality to our historical, moral and political perspective. Postmodernity asks an insidious question which encourages us to reconsider authenticities dating back to the Enlightenment. The Humean and the Hegelian commentaries both eschew the simple answers of equilibrium game theory but do not lose hope, as postmodernity does, that answers exist.

NOTES

* I am indebted to Bob Sugden for introducing me to rational deviance, to Roberto Finelli for some important associations, to Shaun Hargreaves-Heap for the time we

spent arguing about the postmodern condition, and to Joseph Halevi for his dialectical intransigence. Nonetheless, this paper should be blamed entirely on me.

[1] Here I am referring to a specific model of bargaining that has come to dominate the literature: the so-called Nash program. Also note that in game theory, as in most economic analyses, to be rational is to know how to deploy your means effectively in order to achieve your ends. Rationality is exclusively instrumental.

[2] See, for instance, Rubinstein (1982). In his model the distribution will deviate from the 50–50 division to the extent that one player issues her demand *before* the other. As the delay between demands vanishes, the equilibrium outcome tends to be a 50–50 split.

[3] One can, perhaps, accommodate the postmodern view in terms of the computer parallel. Before the theory of chaos, one expected the same algorithm to give identical results if fed the same initial values twice. Since the study of non-linear models has revealed that, because the input *can never be exactly the same twice*, the output may be drastically different. So, why should we expect our two agents to come to the same conclusion? If they espouse slightly different conventions by which to predict the thoughts of others, their train of beliefs may lead them to seriously different conclusions and, thus, disagreement.

[4] See Finelli (1990) who traces the debate on the nature and role of irreconcilable oppositions to the Sophists.

[5] Recall that game theory does exactly this. It starts with simple games, such as the division game described earlier, and assumes solutions for them. Once this stage is over, it then looks at more complex situations (e.g. asymmetric information) and uses the earlier assumptions to obtain explanation. This is what I call the analytic-synthetic method of game theory.

[6] This is a variant of a game that appears quite often in discussions of game theory. See, for instance, Binmore (1987) and Sugden (1989, 1991). For the purpose of easier exposition, I assume that A is female and B is male.

[7] To be precise this is the so-called subgame-perfect Nash equilibrium. A Nash equilibrium is an outcome brought about by strategies which are chosen on the basis of beliefs which are *ex post* confirmed by the outcome. The equilibrium is subgame-perfect if the game comprises more than one stage and such a coordination of strategies and beliefs (i.e. an equilibrium) is achieved not only for the whole game, but, also, in each subgame.

[8] Binmore (1987) criticises over-reliance on backward induction, Sugden (1989) shows that it is possible to have a game theory without this kind of induction provided we are less ambitious and Pettit and Sugden (1990) cement the arguments against the logic in Tables 2 and 3. More recently, Sugden (1991) provides a good summary of the case against backward induction.

[9] Thinking about the possibility of defying the theory that is supposed to govern one's behaviour, is a uniquely human capacity. It is also a capacity that makes the life of the social scientist inordinately demanding. To disallow counterfactuals within a theory (which is what Table 2 does) is to ask for serious trouble since human rationality has the bad habit of instructing agents to ask, 'what if I do not obey the theory's rules?'. In Chapter 6 of Varoufakis (1991), I argue that counterfactual reasoning is, at once, rational *and* incompatible with equilibrium game theory.

[10] The reader may notice that I have made a rather strong assumption concerning what B expects an irrational A to do. Indeed, I assume that an irrational A always does the

opposite of what is good for her. This has allowed us to assume that if $p = 1$ – i.e. if B is convinced that A is irrational – then he expects her to play DOWN at $t = 3$ with certainty. This is, of course, too restrictive. Nonetheless, the main point I am making is not lost if the assumption is relaxed. Suppose, for instance, that an irrational A chooses *as if* by randomisation. Then, at $t = 3$ an irrational A plays UP or DOWN with probability 1/2. In this case, p and q can be re-computed fairly easily and the argument remains intact.

[11] Those familiar with game theory may protest that game theorists recognise the legitimacy of a logic such as that in Table 4 without abandoning game theory's tenets – for example see Kreps et al. (1982). This is correct. However Kreps et al (1982) can only do this after they assume right at the start that agents have some doubt about the rationality of their opponents. Table 4 by contrast does not require such a dilution of the common knowledge of rationality assumption: non-equilibrium strategies are rationalised even when everyone is (at the beginning) absolutely sure that all others are perfectly rational. The ideas in Kreps et al. (1982) become relevant in section 4 in which they help construct a defence of game theoretical orthodoxy.

[12] The reader may ponder the generality of my conclusion in view of the fact that I have focused on a single game. Is it fair to discuss the whole project of game theory on the basis of one example? I think it is. For this is an example that contains a unique Nash equilibrium (subgame perfect) which should, if game theoretical thinking is to be vindicated, produce an unequivocal rational strategy (due to the uniqueness of the equilibrium). If the logic of Table 4 is compatible with full rationality, then we have evidence that the existence of a unique equilibrium does not necessarily tell us what agents will do. Since game theory trades on the thought that it ought to, one example where this is untrue is as good as a thousand.

[13] The *Harsanyi doctrine* occupies a central role in game theory since on it rest a very large number of solutions that would otherwise break down. In the present context, the negative defence draws on it heavily. However there are game theorists who do not accept this doctrine and who, therefore, would not invoke the negative defence. Perhaps they will be inclined to adopt the positive defence of the next section.

[14] Suppose for instance that $q > 1/50$ and A plays DOWN at $t = 1$ but that B sets p at 1/20 and plays UP at $t = 2$. Alternatively, suppose that $q > 1/50$, A plays DOWN at $t = 1$, B sets p equal to 1/8 thus playing DOWN at $t = 2$ and, finally, A plays at $t = 3$. In both these cases one of the two has formed expectations that are proven erroneous ex post.

[15] This is effectively the thesis in Bernheim (1984).

[16] Recall the earlier argument that the equilibrium logic is perfectly legitimate even if not uniquely so. Thus, a player may still choose to be prudent and assume that, since her opponent is equally rational, there is nothing she can do to confuse her.

[17] Table 2, for example, depends on the unique projectibility of traits established at $t = 3$ onto $t = 2$ and $t = 1$.

[18] The game of Table 1 is truly interactive in that what player A does at $t = 1$ depends entirely on what A thinks that B will think if . . . It is in such a game that the enigma of human reasoning becomes pertinent and wrecks the certainty of backward induction. In other cases, where the choice of one player can be made independently of conjectures concerning the actions of another, then of course induction is straightforward. Consider

the following ten dot game. There are ten dots which two players take turns to erase. The first player begins and may erase either one or two dots. Then it is the second player's turn to either erase one or two dots. The player who crosses out the last dot wins. Working backwards, it is clear that the player who plays first has a unique dominant strategy: to erase only one dot at the beginning. In this way, she can be the first to cross out the 4th, 7th and, finally, the 10th dot whatever player B's choices. Backward induction works impeccably in this game because A does not need to consider what B will think if A plays in one way rather than in another. Then, the trait identified at the last stage of the game is uncontentiously projectible to the very first stage when the game commences.

[19] Postmodernity actually rejects the very possibility of a concept. For rhetorical purposes, it may argue that analogies are often mistaken for concepts, in order to demonstrate the vacuousness of concepts.

[20] The reader who would like to leave open the possibility that an irrational player acts in an unpredictable manner will protest that this is too stringent an assumption. However, the analysis will not change significantly if we envision an irrational player A as someone who chooses between UP and DOWN as if by randomisation. Footnote 10 applies here with equal force.

[21] I assume that A believes B to be rational with probability $1 - w = 1$. The model can be easily extended to allow for two-sided uncertainty concerning rationality, i.e. letting $w > 0$.

[22] For the updating mechanism to make intuitive sense, $1 - r > s$, i.e. the probability that A will adopt some logic different to that of Table 2 (RBI) if rational and thus choose the deviant strategy must exceed the probability that an irrational A will choose the deviant strategy. This is very sensible since otherwise there would be no reason for B to believe that DOWN at $t = 1$ enhances the prospects that A is irrational.

[23] Naturally, part (k) of Table 6 is tantamount to the condition $p_1 > 1/21$.

[24] Binmore (1987, 1988) has also voiced concern about the undiscriminating use of Bayes' rule.

[25]. There is an interesting parallel here with Foucault's (1967) critique of 'modernity's monologue'. Foucault claims that before the triumph of modernity, there used to be a dialogue between rationality and madness. Later, this dialogue broke down and left us with a monologue of rationality on madness. And yet, he goes on, there are dimensions of sense in madness that are missing in what we tend to think of as Reason, or to put it differently, there is a great deal of Reason in madness. Any attempt to evict madness altogether in order to procure pure Reason is, therefore, ill-conceived. The reader who is so inclined may interpret the assumption that there exists a uniquely rational estimate of s as a technical manifestation of illegitimate attempts to cement this monologue.

[26] A player's conceptualisation of her opponent's conjectures is, in itself, a theory. To argue that one attaches, via induction, probabilities to different such theories and, in addition, to insist that these probabilities are common property, is philosophically absurd. Peirce (1932) draws the important distinction between the probability *of* a hypothesis and the probability *derived from* a hypothesis. He writes:

> It may be conceived, and often is conceived, that induction lends a probability to its conclusion. Now that is not the way in which induction leads to the truth. It lends no definite probability to its conclusion. It is nonsense to talk of the probability of a law, as if we could pick universes out of a grab bag and find in what proportion

> of them the law held good . . . What induction does . . . is infinitely more to the purpose.

[27] The reason is that if A goes DOWN at $t = 1$, then Equation (1) will update B's probabilistic assessment that A is irrational to $p_2 = 1/11$. This posterior belief makes B indifferent between UP and DOWN at $t = 2$. Thus, A anticipates that DOWN at $t = 1$ will make B randomise at $t = 2$, a thought that makes her unsure as to whether she ought to play DOWN at $t = 2$. Consequently, she also randomises at $t = 1$.

[28] See Skyrms (1990) for a discussion of deliberational disequilibrium.

[29] Table 4 has presented this critique of equilibrium theory implicitly. Let $p_1 = 0$. Then, if player A played DOWN at $t = 1$, Equation (1) cannot be defined: an event occurred that B had attached a zero probability to. So, what should B do in such a situation? According to the equilibrium story, there is no answer. Can we speculate that, in the absence of advice by the theory, player B may still revise p upwards (i.e. $p_2 > 0$). If A expects this to happen (and there is no reason why she should not), then she may rationally choose DOWN at $t = 1$. Of course, there can be no equilibrium account of what has happened. Therefore, equilibrium theory is inferior to the account of table 4 because rational agents may have an incentive to violate it.

[30] Nietzsche wrote: "What therefore, is truth? A mobile army of metaphors, metonymies, anthropomorphisms; truths are illusions of which one has forgotten they are illusions . . . coins which have their obverse effaced and now are no longer of account as coins but merely as metal" *On Truth and Falsity on their ultramoral Sense* in Levy (1964). However, Heidegger successfully exposed holes in his arguments by demonstrating that Nietzsche needs truth as a concept in order to argue against its meaning. Interestingly, this is also a problem for Heidegger. Finelli (1990) claims that Being is denied by Heidegger and his contemporary postmodernist followers, but that in their philosophy it returns to determine human reality through its loss and emptiness.

[31] Sugden (1991) illustrates this point in the context of (i) a critique of Savage's expected utility theory and (ii) the theory of games.

[32] Gerhard Adler writes: "The enigma of creativeness rooted in the irrational, indefinable matrix of man's timeless psyche has held eternal fascination for him and has helped produce the most memorable justification of his status as man" – see the *Foreword* in Kirsch (1966).

[33] The social anthropologist Levi-Strauss (1966) defines analytical Reason as the type of logic that develops when humans try to understand natural phenomena of a low order (e.g. hydrodynamics as opposed to the concept of time). He thinks that such logic is frustrated when it is called upon to explain social phenomena. The result of this failure is a new kind of Reason which, in Hegelian fashion, he terms dialectical. ". . . [D]ialectical reason thus covers the perpetual efforts analytical reason must take to reform itself if it aspires to account for language, society and thoughts; and the distinction between the two forms of reason in my view lies on the temporary gap separating analytical reason from the understanding of life. Sartre calls analytical reason reason in repose; I call the same reason dialectical when it is roused by action, tensed by the effort to transcent itself."

[34] Of course, postmodernity is eager to attack Hegel in the same way that it disparaged Hume. The contradiction on which Hegel bases the sublation of Reason is seen as both unresolvable and as unreal. It is unresolvable because Reason is meaningless and,

therefore, hardly capable of improving itself. It is unreal because when we talk of *the* contradiction, we fall victims to the inferiority of our language. The latter is forced, through its imprecision, to contrive false categories (such as Reason and Unreason) when, in reality, *the* contradiction is, like truth in Nietzsche, an illusion that we have forgotten that it is an illusion.

References

Aristotle: 1987, *Nicomachean Ethics*, transl. by J. Welson, Prometheus, New York.

Bernheim, D.: 1984, 'Rationalisable Strategic Behaviour', *Econometrica* **52**, 1007–28.

Binmore, K.: 1987, 1988, 'Modeling Rational Players: Parts I and II', *Economics and Philosophy* **3**, 179–214 and **4**, 9–55.

Derrida, J.: 1978, *Writing and Difference*, Routledge and Kegan Paul, London.

Finelli, R.: 1990, 'Production of Commodities and Production of Images: Reflection on Modernism and Postmodernism', mimeo, Faculty of Philosophy, University of Rome.

Foucault, M.: 1967, *Madness and Civilisation*, Tavistock, London.

Harsanyi, J.: 1973, 'Games with Randomly Disturbed Payoffs: a New Rationale for Mixed Strategies', *International Journal of Game Theory* **2**, 1–23.

Hegel, G. W. F.: 1931, *The Phenomenology of Mind*, translated by J. Baillie, London.

Kirsch, J.: 1966, *Shakespeare's Royal Self*, Putnam and Sons, New York.

Kreps, D., P. Milgrom, J. Roberts, and R. Wilson: 1982, 'Rational Cooperation in the Finitely Repeated Prisoner's Dilemma', *Journal of Economic Theory* **27**, 245–52.

Levi-Strauss, C.: 1966, *The Savage Mind*, Weidenfeld and Nicholson, London.

Lyotard, J.-F.: 1984, *The Postmodern Condition: A Report on Knowledge*, Manchester University Press, Manchester.

Norris, C.: 1985, *The Contest of Faculties: Philosophy and Theory After Deconstruction*, Meuthen, London.

Nietzsche, F.: *On Truth and Falsity in their Ultramoral Sense*, in Levy, O.: 1964, *The Complete Works of Friedrich Nietzsche*, New York.

Peirce, C. S.: 1932, *Collected Papers Vol. 2*, Harvard University Press, Cambridge Mass.

Pettit, F. and R. Sugden: 1989, 'The Paradox of Backward Induction', *Journal of Philosophy*, **LXXXVI**, 169–82.

Skyrms, B.: 1990, *The Dynamics of Rational Deliberation*, Harvard University Press, Cambridge Mass.

Sugden, R.: 1989, 'Game Theory without Backward Induction', mimeo, University of East Anglia.

Sugden, R.: 1989b, 'Spontaneous Order', *Journal of Economic Perspectives*, **3**, 118–25.

Sugden, R.: 1991, 'Rational Choice: A Survey of Contributions from Economics and Philosophy', *The Economic Journal* **101**, 751–85.

Varoufakis, Y.: 1991, *Rational Conflict*, Basil Blackwell, Oxford.

Manuscript submitted 27 February 1992
Final version received 19 June 1992

Department of Economics
University of Sydney
Sydney 2006
Australia