International Institute for
Applied Systems Analysis
Schlossplatz 1
A-2361 Laxenburg, Austria

Tel: +43 2236 807 342
Fax: +43 2236 71313
E-mail: publications@iiasa.ac.at
Web: www.iiasa.ac.at

**Interim Report IR-09-091**

# The bounds of reason and the joys of virtue

Karl Sigmund (ksigmund@iiasa.ac.at)

**Approved by**

Ulf Dieckmann
Leader, Evolution and Ecology Program

June 2010

***Interim Reports*** on work of the International Institute for Applied Systems Analysis receive only limited review. Views or opinions expressed herein do not necessarily represent those of the Institute, its National Member Organizations, or other organizations supporting the work.

Karl Sigmund: Book Review (for the American Scientist) of Herbert Gintis, *The Bounds of Reason: Game Theory and the Unification of the Behavioural Sciences,* Princeton University Press (2009)

# The Bounds of Reason and the Joys of Virtue

Humans are social animals, and so were their ancestors, for millions of years before the first camp-fires lighted the night. But the mathematics of social interactions is a recent acquisition. It originated in the 'forties of the previous century, with a book originally titled 'The Theory of Rational Behavior'. Its authors, the mathematician John von Neumann and the economist Oskar Morgenstern, eventually re-named it 'Game Theory and Economic Behavior', a fortunate move. However, the basic assumption in this book remained unaffected: it postulated, as did all follow-up texts on game theory for generations, that players are rational.

Three decades later, game theory experienced a new lease on life through the work of biologists William D Hamilton and John Maynard Smith. It was applied to analyse biological interactions, such as inner-specific fights or parental investments. This new 'evolutionary game theory' was no longer based on axioms of rationality. In fact, one of the pillars of classical game theory, Anatol Rapaport, characterized it as 'game theory without rationality'. Herbert Gintis was among the first economists attracted by that new field, and when, ten years ago, I wrote a review of his text-book on 'Game Theory Evolving', I described it as 'testimony of the conversion of an economist'. Gintis has not recanted in the meantime: indeed, a second edition of that book just appeared. But a companion volume, entitled 'Bounds of Reason', shows that he certainly has not forgotten his upbringing in the orthodox vein.

There is, of course, no contradiction between the two game theories. As with any mathematical theories, their aim is to rigorously derive the consequences of well-defined assumptions which are taken as granted. Just as there are geometries using the parallel axiom, and others that do not, so there are game theories using rationality axioms and others that do not. Mathematically, they are of equal respectability. It is only when they are applied to the real world that tensions arise.

In his preface, Gintis compares the history of physics with that of economy. Physical theories have regularly stumbled against experiments revealing ‘anomalies’ which falsified them and led to their replacement by other theories. Compared with such an intensive dialogue between theory and experiment, most textbooks of classical economy remained singularly fact-free. This has changed recently: experimental economics, and in particular behavioral game theory, has flourished in the last decades, and ‘The Bounds of Reason’ contains an impressive catalogue of empirical findings. In the light of conventional rationality assumptions, many of them appear as ‘anomalies’.  They are based on intriguing thought experiments exploring the many facets of decision making.

The most venerable is the Prisoner’s Dilemma. Imagine two players having simultaneously to decide whether to send a gift to the co-player. Sending it costs the donor 5 dollars, and provides the recipient with 15 dollars. If both send their gift, they earn 10 dollars each. But whatever the co-player does, a player earns more by not sending the gift. Two rational agents who want to maximise their income will end up  getting nothing. The same happens if one player decides after being informed of the other’s decision. The Trust game is closely related. In this game, a coin toss decides who of the two players is the Investor, and who the Trustee. The Investor can send a gift to the Trustee, as before. The Trustee can then return part of the sum to the Investor. Thus both players can profit from the exchange. But rational players are stuck in a dead end: no gift, no return. Real players frequently overcome the stalemate. They trust, and honour trust, and mutually profit from their interaction.

Another example of some entertainment value is the Traveller’s Dilemma. Two persons have to separately file a claim between 10 and 100 dollars. The rules prescribe that if they both claim the same amount, both will get it; but if they claim separate amounts, the lower will be judged to be valid and paid out. As an incentive, the more modest claimant receives two dollars more, and the other player correspondingly two dollars less. Two innocent players would immediately claim 100, and get it. But a smart player bent on optimizing might hope to get one dollar more, by merely claiming 99. Fearing that the co-player might think along these same lines, one should not claim 100, because this would yield only 98. But if both claim 99, that's what they get: better to outsmart the other by claiming 98, and pocket 100…And from then on, the argument, an instance of what is termed 'backward induction', leads unflinchingly all the way down to the minimum claim of 10 dollars. Both players are caught in a trap: asking for the minimum is the only solution which is consistent in the sense that no player can gain by unilaterally deviating from it.

Needless to say, real people rarely reach this solution, in experiments: they may take a few steps down the ladder, but then stop well above 90.

Another classical example is known as Ellsberg paradox. Consider an urn containing 30 balls which can be red, black or white. All you know is that 10 balls are red. The number of black or white balls is unknown. You now are offered to bet on a colour: if it is drawn, you will receive 100 dollars; else, nothing. Most people bet on red, and this seems hardly surprising. But if asked to bet on any combination of two colours, instead, most people prefer to bet on (black, white), and this not consistent with their former choice. In another vein, most people prefer a price of 10 dollars, with a 1 dollar discount, to a price of 8 dollars with a 1 dollar surcharge. Such framing effects are ubiquitous.

Under the relentless impact of these empirical findings, the assumption of rationality has been modified in various ways. From its glorious state of an unreachable idealization, it has been downgraded to 'bounded rationality'. This view accepts that perfect optimization is, in general, beyond human reach. But bounded rationality means different things to different people, depending on which parts of the ideal they want to salvage. This has given rise to interesting theoretical investigations into the logical dependencies of diverse blends. Nevertheless, many doubt that it helps to account for the empirical findings. The situation is reminiscent of the quandaries of early astronomers, who discovered, to their dismay, that planets did not move in perfect circles as presumed. They came to terms with this ‘bounded circularity’ by describing the orbits as epicycles – circles within circles, and so on – and ingeniously deriving fairly decent approximations, at the cost of an explosion in sophistication and complexity.

The version of rational agents proposed by Herbert Gintis gets rid of the assumption that players enjoy ‘common knowledge of rationality’. Thus players are all rational but need not believe that all of them are rational. This helps indeed to avoid the trap of backward induction. On the other hand, it does not explain why people are so sensitive to framing, or so inept at lotteries and bets. Gintis, who claims that ‘the bounds of reason are not the irrational, but the social’, dismisses anomalies involving just one single person as ‘performance errors’. Such a fix can explain everything and nothing, and therefore threatens to weaken his cause.

For Gintis champions a cause, one which can be wholeheartedly subscribed: namely to promote game theory as an indispensable instrument in modelling human behavior. He rightly points out the wide discrepancies between the approaches in separate branches of the behavioural sciences: for example, economics, sociology, anthropology and social psychology use vastly different premises in studying social behaviour and organization. Such academic traditions (based, in Gintis’ term, on ‘virtually impassible feudal organizations’) are not likely to vanish overnight, but behavioral game theory can offer a tool for them all. In fact, we see  this happening already. To give a few examples: (a) the behavior in Trust games and similar games of cooperation is a superb ‘microsocial’ indicator for what sociologists call the Rule of Law (the general respect for rules and institutions in a society); (b) game theoretical models help in explaining the adaptive value of humankind’s nearly ubiquitous belief in supernatural agents; (c) ingenious experiments uncover our  often subliminal concern with being watched, which can be triggered by the mere image of an eye and greatly boosts cooperative behaviour; (d) the widely varying gift-giving traditions in small-scale societies, with all their attending complexities concerning obligations and status, are dissected by means of simple economic experiments, etc.

The academic tribes, however, will hesitate in accepting the gift of game theory from economists if they are told that it comes with the rational actor model. Not everyone wants to shoulder the obligations that go with it. No doubt that humans have developed the faculty of reasoning to a unique degree; but our decisions are also guided by other factors, such as habits and customs, passions, emotions, and 'animal spirits' (to use the expression of economist John Maynard Keynes). Many actions do not fall under the heading of rational behavior, as commonly understood (although it must be admitted that modern economists' definitions of rationality are as far removed from the every-day use of that word, as modern theology's concepts of Divinity from the average lay-person's idea of the Good Lord).

Psychologists, for instance, analyse decisions in terms of (often unconscious) cues and heuristics, and are not likely to switch to the paradigm of Beliefs, Preferences, Constraints and Expected Utilities which underlies the rational actor model. Why should they? In evolutionary game theory, they can enjoy the full panoply of behavioural experiments without the restraints imposed by the loitering presence of the rational actor. Strategies (i.e., programs of behaviour) need not be the product of rational decisions. They can be copied, for instance, through learning or inheritance.

Interestingly, Gintis stresses right from the start that game theory is central to understanding the dynamics of all sorts of life forms, and touts it as 'the universal lexicon of life'. Since it applies to primates, birds, lizards, plants and even bacteria, it must be able to do without the rational actor model. Needless to add that in applications to humans, we must take account of our very special cognitive and communicative facilities. But in the 'Bounds of Reason', Gintis allows only a marginal role for evolutionary game theory. Surprisingly, he hardly mentions the pioneering work of Trivers, Hamilton, or Axelrod. This is a pity, since what these authors have to say about human nature had a deep impact on most behavioral sciences. In the tradition of Adam Smith, evolutionary biologists explain cooperative behavior by long-term self-interest ultimately grounded on reproductive success. In particular, it can be advantageous to forfeit an immediate material gain if this increases one's reputation, and thus promises to confer a higher value in the market for collaborators. The corresponding cost-benefit calculation need not be conscious, let alone rational; it can be mediated by heuristics based on emotional responses such as shame, sympathy or anger, and possibly rationalized after the act.

Gintis embraces another approach, and explains cooperation by a human preference for what he terms character virtues (such as honesty, trustworthiness or fairness). But every behavior can be interpreted as a preference for some virtue. Short of providing an ultimate reason for the preference (e.g., that it promotes long-term self-interest), this approach has as little explanatory power as Gintis' statement that 'the increased incidence of prosocial behaviour is precisely what permits humans to cooperate effectively in groups'. Likewise, he postulates a genetic predisposition to follow social norms even when it is costly to do so. But neither the evolution of the predisposition nor the emergence of norms are explained. Many game theorists have attempted to describe collective phenomena, and in particular social norms or social institutions, by the actions of the individuals involved.  Gintis, however, rejects methodological individualism – the basis for such an approach – on the ground that it is incompatible with the characteristics of rational agents. But incompatibility cuts both ways.

To resume, 'The Bounds of Reason' appear as two books in one. One part develops an epistemic theory of the rational actor as an alternative to what is provided by classical game theory, and the other part is a spirited plea to use behavioral game theory as unifying tool in all behavioral sciences. Both objectives are highly valuable, but combining them creates friction. Friction creates heat, and Herbert Gintis, who thrives gleefully on controversial issues, may have enjoyed the prospect of heated discussions.