DRAMA THEORY AND METAGAME ANALYSIS

Nigel Howard
ISCO Ltd, 12 Chesham Road, Brighton BN2 1NB, UK

Keywords: drama theory, confrontation analysis, metagame, game theory, conflict, resolution, dilemmas, rationality, rational choice, preference change

Contents

Summary

An instrumentally rational approach to conflict generates game-theoretic “dilemmas”. Mainstream game theorists have decided to live with these; others see them as casting doubt on whether humans can or should be rational. Metagame theory attempted to overcome the dilemmas by formally representing a great depth of mutual understanding between players. Dilemmas of *agreement* (typified by prisoner’s dilemma) are overcome in this way; dilemmas of *disagreement* (typified by chicken) are not. Meanwhile, metagame theory, being essentially non-quantitative, gave birth to an applied technique – metagame analysis – which has helped to resolve many real-world conflicts. Experience with seeing how dilemmas were dealt with in the real world gave birth to drama theory. This proposes that at a “moment of truth” – a point when players have communicated what they see as their final “positions” – dilemmas cause positive and negative emotions which jolt players out of their fixed ways of seeing the situation and enable them to redefine it in such a way as to eliminate dilemmas. When all

dilemmas have been eliminated, it is proved that all players are taking the same position and can trust each other to implement it. In this way the game is transformed at successive moments of truth until a totally satisfactory type of solution becomes possible and is unanimously selected by all players. Thus drama theory takes into account the effects of both emotion and rational debate – factors omitted from game theory. Instead of taking the game as fixed, drama theory allows players' preferences and perceived opportunities to vary as they are moved by emotion to redefine the game itself, as well as their positions in it. Thus drama theory departs from instrumental rationality. At the same time, it takes a strong interest in game models and game-theoretic analysis, seeing them as generating the dilemmas that motivate dramatic transformations of the game itself.

1. Dilemmas generated by a rational approach to conflict

Conflict (mutual attempts by people to harm or frustrate each other) can be explained in two ways: as automatically resulting from causes – genetic or sociological – that operate upon humans; or as "rational" – i.e., as a result foreseen by parties consciously pursuing their aims.

A classic example of the first, "causal" approach is Richardson's model of the arms race (L.F. Richardson, *Statistics of deadly quarrels*, Quadrangle Books, Chicago, 1960). This is a set of differential equations in which each side's increase in armaments is positively affected by the other's actual level of armaments and negatively affected by its own. Conflict, in the sense of actual use of armaments to inflict harm, occurs when the process becomes unstable through positive reinforcement so that the arms race "explodes".

Richardson's model is organizational; it looks at how a human organization reacts to its environment. Individual people are mere elements of the organization, bound to obey its laws. Other causal models are psychological. Usually, they assume that conflict results from people becoming "aggressive". Aggressiveness is seen either as a basic drive that persists until satisfied or as a learned response, reinforced or discouraged by effects experienced from it.

Though some causal models seem convincing, it is significant that all are incomplete in that they also leave room for rationality. Richardson sees the arms race as likely to take place *unless* we take steps to prevent it; but taking steps would be a rational decision, with the aim of avoiding war. Similarly, psychological theories point out ways in which aggressiveness, and hence conflict, can be reduced – egg, by changes in education or child-rearing or by media censorship. Making such changes would again be a rational step, taken to reduce conflict.

Thus all approaches leave room for rational decisions either by organizations – such as a society or nation as a whole – or by individuals. It is, in fact, typical of the rational approach to decision-making that rationality (i.e., choosing behavior in conscious, deliberate pursuit of objectives) may be attributed either to individuals or organizations.
What happens, then, when we analyze conflict on the assumption that people, as individuals or organizations, always behave rationally, given their beliefs about the situation they are in? Game theory is the discipline that sets out to answer this question.

Its basic model supposes a number of parties (individuals or organizations) each of which must choose between a number of options, to be implemented all at once or sequentially. Each party has its own preferences as to the options that itself and others will choose, and knows not only its own options and preferences, but others' also -- though its knowledge may be no more than probabilistic. Within this given framework, each, being rational, "optimizes" – i.e., chooses so as to get the most preferred outcome for itself, given its beliefs about how others will choose. This definition of rationality implies that rational players must, if their beliefs turn out to be correct, find themselves at a so-called *equilibrium point* – a point where each is optimizing against the others' choices.

Alarmingly, this model generates conceptual difficulties, called "dilemmas". These are so disconcerting as to cast doubt on the validity of the whole rational approach to decision-making.

The best-known occurs in the so-called "prisoner's dilemma" -- a game where a sheriff holding two prisoners gives each of them a choice to *confess* or *not confess*. If one alone confesses, it is freed while the other gets a maximum sentence. If neither confesses, both get a light sentence. If both confess, both get a heavier but less than maximum sentence. In this situation, the only rational choice for a sentence-minimizing prisoner is to confess, since this gives it a better outcome, whatever the other does. Hence *both are worse off if both are rational than if both are irrational*. Specifically, each gets a lighter sentence if both try to maximize their sentences than if both try to minimize them.

Another dilemma occurs in the game of "chicken", in which two teenagers drive cars at each other. The first to swerve is shamed while the other triumphs – but if neither swerves, both die. This has two possible equilibrium solutions, one better for one player, the other for the other; the solution each favors, and implicitly proposes, is that the other swerve first. Which solution will be chosen? The answer: *by being rational a player allows the other (if prepared to be irrational) to obtain the solution that is better for itself -- and worse for the rational player.* Again, irrationality pays.

2. Reaction to the dilemmas; metagame theory

Far from being mere curiosities, game-theoretic dilemmas occur throughout social life. They are typical of situations in which interdependent parties have differing objectives. Some see this as requiring a modification of the assumption that people can or should be rational in the above strict, game-theoretic sense. Most game theorists, at least since the sixties, disagree. At that time the dilemmas were taken seriously; since, the mainstream reaction has been to formally ignore them (while relishing the piquancy they bring into theorizing), and continue to analyze game situations on the assumption of strict rationality.

Reacting against this "rationalist" tendency, metagame theory attempted to resolve dilemmas by requiring more of rationality than that each player should optimize given its *beliefs* about the others' choices. The new requirement was that each should be able to *know* the others' choices, and *know* how the other would choose to react to such knowledge, and *know* each other's reactions to such reactions, and so on. Such depth of mutual understanding might come about through good communication. The question was: what would it achieve?

This was answered by defining an infinite set of "metagames" relative to a given game. Von Neumann and Morgenstern, the creators of game theory, had tackled the basic problem of game theory (*how to choose optimally when my optimal choice depends on yours, and* vice versa) by defining the "minorant" and "majorant" games of a given two-person game; these were obtained by supposing that player 1 (respectively 2) could choose *after* 2 (respectively 1), and therefore in knowledge of 2's (1's) choice. A rational solution to a game, they averred, would correspond to an equilibrium point of both its minorant and majorant games. Metagame theory goes further and looks at the minorant of the majorant, the majorant of the minorant, and so on. Thus it defines, for each player i in a given game G, the "metagame" iG; this is the game in which player i chooses in knowledge of the choices of all the others. It is then possible to form all the games $i_1 \ldots i_m G$, where $i_1 \ldots i_m$ is any string of players' names. This yields an infinite tree of metagames, modeling the possibility that players may be able to gauge each others' intentions, and their intentions given knowledge of each others' intentions, to any depth. A further development of metagame theory enlarged the metagame tree further by allowing for the possibility of partial, incomplete knowledge of other players' intentions.

The result is that the dilemmas of cooperation such as in prisoner's dilemma are "solved" without the need to go far into the metagame tree. Players who can know, not only what each other will do (a 1^{st}-level prediction), but how each other will react to such knowledge (a 2^{nd}-level prediction) are strictly rational to cooperate and trust each other, since such cooperation corresponds to a rational equilibrium in the metagame tree.

It is worth noting that the results of metagame theory parallel those of another approach -- *repeated game* theory. This assumes a game with additive, time-discounted utilities that is repeated between the same players an infinite number of times. Though unrealistic compared to the metagame approach, this set of assumptions has proved more acceptable to mainstream theorists. Under it, dilemmas of cooperation again disappear.

3. Dilemmas of agreement and disagreement; metagame analysis

Metagames (or repeated games) did not, however, "solve" all dilemmas.

Dilemmas are broadly of two kinds. The prisoner's dilemma arises despite the fact that players *agree in proposing the same solution*. Their problem is that they cannot trust each other to implement this solution – and indeed, cannot trust themselves to do so – since *provided one implements it, the other is better off not doing so*.

Dilemmas of agreement such as this are "solved" by metagames and repeated games. Getting trust is not, however, the only important problem of social life – though the attention paid to prisoner's dilemma by social scientists and evolutionary theorists sometimes seems to suggest it is. Equally important are dilemmas like that in chicken.

Here the problem is not that players cannot trust each other to implement an agreed solution; it is that *each is proposing a different solution*. If they agreed on any one solution, it would be stable. But they do not.

This presents them with a different set of dilemmas we call *dilemmas of disagreement*. Of course, the two kinds of dilemma can occur in the same game. Players might propose different solutions, and it might also be the case that if they agreed on a solution, they would encounter a problem of mutual trust. But the dilemmas involved are different.

Despite its failure to solve dilemmas of disagreement, metagame theory, being an approach which does not require quantitative utilities, led to an applied technique for analyzing real problems called *metagame analysis* or *options analysis*. It was first used under contract to the US Arms Control and Disarmament Agency to analyze negotiations leading to the first SALT agreement.

The basic technique finds the "metaequilibria" (points corresponding to equilibria in the metagame tree) of real-world political or economic conflicts. Metagame theory proves that these are the points from which each deviation by a subset of players is punishable by a reaction of the complementary subset. This is precisely the definition of the *core* of a game, so that metagame analysis is, in effect, a practical, non-quantitative way of first exploring what game theorists call the *strong equilibria* of a game, then exploring its *core*.

How does it work? Players' strategy-sets are specified by assigning each player a number of yes-no options, later called "cards"; the player chooses a strategy by deciding which cards to play, and which not – though its choice in relation to a particular card may be constrained as a so-called "consequence" of other cards being played or not played. The possibly sequential nature of strategy implementation can be captured by supposing that opportunities to play cards come in a certain order.

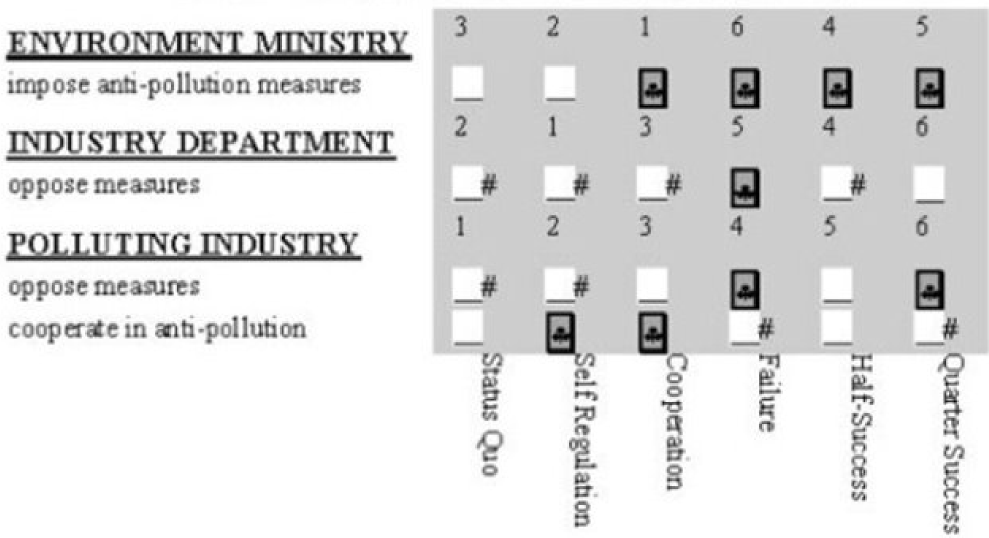

Figure 1: Card-table model of an anti-pollution problem

A game is then represented by a tableau, or "card table", like that in figure 1, which is taken from a study of anti-pollution measures carried out for a European government. Here each column represents an outcome of the game. In each column, the playing of a card is represented by a picture of a card, its non-playing by a white space. Where a white space has a hash mark ("#") beside it, the non-playing of that card is an assumed "consequence" of the playing or non-playing of other cards in the same column; for example, in column 1 the simple assumption is made that if anti-pollution measures are not imposed, they cannot be opposed.

Outcomes (columns) in a card table are interpreted as different possible "futures" of the situation. Player's preference rankings of these "futures" are indicated by the numbers opposite players' names, with 1 being assigned to the most preferred column, 6 to the least preferred.

Any future *s* can then be analyzed to show, first, whether there is any subset of players that can, by altering just their own selection of cards, move to a future they all prefer to *s*. If a subset *G* of players has such a so-called "unilateral improvement" from *s*, the next step in the analysis is to ask whether the complementary subset of players can, by altering their own selection of cards, respond to the unilateral improvement in a way that makes some player in *G* worse off (or at least, no better off) than at *s.* If so, then the reaction of the players not in *G* is called a "sanction" against *G*'s unilateral improvement.

In this way the analysis determines whether, and if so, how, any particular future *s* can be "stable"; it can be stable if every credible unilateral improvement is deterrable by a credible sanction. Note that basic metagame analysis does not determine whether particular improvements or sanctions are, or can be made, credible. Other methods, including the user's judgment, must be used to decide this.

4. From metagame analysis to drama theory

Metagame analysis helped business and governmental users to resolve conflicts. But experience with using it on real problems was found to require more than setting down all the options ("cards") held by each player and, having done so, analyzing the stability of the futures thus defined.

First, there had to be some way of deciding which possible actions, out of a virtual infinity, were relevant "cards" for each player. Second, even a small number of cards generated astronomical numbers of possible futures; it was necessary to select which ones to analyze.

These are technical problems. A more basic problem was that, having analyzed the stability of certain futures under certain assumptions, dilemmas were exposed. Players saw and felt these. They might then become emotional and solve the dilemmas by *redefining the game or their objectives within it.* A new analysis, using different assumptions, was then required.

Such behavior is quite un-game-theoretic. Nevertheless, this phenomenon of dilemmas causing both *emotion* and *attempts at dilemma-elimination*, seemed to give a clue both as to how players were likely to behave and as to how a conflict was resolved. Eventually, it was theorized that while dilemmas existed, there would be tension between players, causing emotion of two kinds: *positive* emotion (love, goodwill), which had the function of *making promises credible* and thereby eliminating dilemmas of agreement; and *negative* emotion (anger, resentment), with the function of *making threats credible* and thereby eliminating dilemmas of disagreement. These emotions enabled players to change their beliefs and values in such a way as to rationalize their redefinition of the situation. When, finally, no dilemmas were left, the conflict was resolved.

Drama theory, initiated in the early nineties, formalized and developed these ideas. This approach avowedly looks at a pre-play period of communication between players called a *confrontation*. In this period, prior to finally deciding which strategies to implement, players attempt to define not only the game but also their "positions" within it in such a way as to resolve their interdependent decision problem. If they succeed, the game they then play has a trivial, agreed solution – and no dilemmas. If they fail, dilemmas generate emotions and rationalizations through which they *redefine* the game and/or their positions. This process tends to continue till all dilemmas are eliminated – at which point it is proved that all players take the same position and can trust each other to carry it out.

An important drama-theoretic innovation is that the object of study – the object changed via emotion and rationalizations – is not just a game. It is a game together with *positions*. Each player is assumed to have taken both a positive position (stating the outcome it wants everyone to implement) and a "fallback" position (stating the unilateral strategy it proposes to implement if its position is not accepted). Dilemmas are defined with reference to this object -- a game-with-positions – not with reference to a game on its own. Dilemmas exist in prisoner's dilemma and chicken only because it seems obvious what positions each player will take in these games; in general, however, positions need to be specified before dilemmas can be analyzed. For example, in Figure 1: Card-table model of an anti-pollution problem, the position of the Environment Ministry was Cooperation (column 3); that of the Industry Department was Self-Regulation (column 2); and that of the polluting industry was Status Quo (column 1). The "fallback" outcome (expected if every player carried out its fallback position) was Failure (column 4).

By assuming this added structure in the object studied, drama theory overcomes the two "technical" problems of *selecting relevant options* from among a virtual infinity of possibilities and *selecting futures to analyze* from the large number generated. Relevant options are those necessary to define players' positions. Futures requiring analysis are the positions themselves – both players' positive positions and the future generated by their fallback positions.

The term "drama" is used in place of "game" because in the assumed pre-play period of attempted conflict resolution the phenomena of drama, as distinct from game-playing, are found. Actors use emotion and rational debate to try to change their own and each others' beliefs and values. Means such as these, not covered by game theory, are necessary because it's assumed that players cannot simply *choose* how they will define the game. Their definition is based upon their honest beliefs about the opportunities open to them and their effective value systems. Thus it is based upon characteristics of the players and the world they inhabit that are changed by interactions involving emotion, reason, exploration, debate and the exchange of threats and promises, rather than by calculations based upon instrumental rationality.

By such means people and organizations undergo deep change. This is the stuff of drama.

-
-
-

Bibliography

Fraser, N.M. & K.W. Hipel. (1984). Conflict Analysis: Models and Resolutions. North-Holland. [A technique for analyzing conflicts that takes off from metagame analysis and develops it with certain further assumptios.]

Howard, N. (1971). Paradoxes of rationality: theory of metagames and political behavior. Cambridge, MA: M.I.T. Press. [The main source for metagame theory; metagame analysis is also introduced.]

Howard, N. (1989). The Manager as Politician and General. In J. Rosenhead (ed.), Rational Analysis for a Problematic World. Chichester, UK: Wiley. [An advanced exposition of metagame analysis.]

Howard, N. (1996). Negotiation as drama: how 'games' become dramatic. International Negotiation 1, 125-152. [A view of drama theory as a technique for analysing negotiations.]

Howard, N. (1998). n-Person 'Soft' Games. Journal of Operational Research Society 49, 144-150. [Proves the fundamental theorems of drama theory.]

Howard, N. (1999) Confrontation Analysis: How to Win Operations Other Than War. CCRP Publications, Penatagon, Washington DC. [Applies confrontation analysis – a technique for applying drama theory – to peace-keeping and the problems of defense in the post-Cold War world.]

Howard, N., P. Bennett, J. Bryant, & M. Bradley. (1992). Manifesto for a theory of drama and irrational choice, Journal of Operational Research Society 44(1): 99-103. [The first introduction to drama theory, setting out a program of research and applications.]

Radford, K.J. (1977). Complex Decision Problems. Reston, VA. [A common-sense approach to use of metagame analysis, allying it with other decision-support techniques.]

Rapoport, A. (1960). Fights, Games and Debates. New York: Harper. [A classic overview of the signficiance of game theory and the dilemmas to which it gives rise.]