Group Decision and Negotiation, 3:207–235 (1994)

Drama Theory and Its Relation to Game Theory. Part 2: Formal Model of the Resolution Process

NIGEL HOWARD
Nigel Howard Systems, 10 Bloomfield Road, Birmingham B13 9BY, UK)

Abstract

In a drama, characters' preferences and options change under the pressure of pre-play negotiations. Thus they undergo change and development. A formal model of dramatic transformation is presented that shows how the core of a drama is transformed by the interaction among the characters into a strict, strong equilibrium to which they all aspire. The process is seen to be driven by actors' reactions to various "paradoxes of rationality."

Key words: drama theory, game theory, preference change, emotion, soft games

1. Introduction

Part 1 of this article (Howard 1994) gave a general outline and discussion of drama theory, together with an applied example. This second part gives a formal model and a more detailed, formal discussion of the dramatic resolution process.

The basic idea of drama theory, explained in part 1, and also in Howard et al. (1993), is that various game-theoretic paradoxes undermine the very concept of rational choice, defined simply as *choosing A rather than B when A is both preferred and available*. These paradoxes are, of course, well known. The innovation of drama theory is to posit that characters who confront such paradoxes tend to feel various kinds of emotion, which help them to avoid a "breakdown of rationality" by providing the energy needed for them to "reframe" their situation. Reframing may consist of perceiving new, hitherto unperceived options for themselves or others, or it may consist of characters changing their preferences.

Thus, the fundamental difference between drama theory and game theory is that a drama allows for the possibility of the game itself changing even though the environment remains informationally closed; that is, it considers the possibility of endogenous changes, arising from interactions within the game itself. This difference seems so important that a new metaphor and corresponding vocabulary seem appropriate. The metaphor of *drama* is thus used to encompass the idea of an interaction in the course of which characters change, develop, and perceive things in a new way.

In reframing their situation at the climax, characters in a drama develop their personalities and systems of values. They also develop as a community, constructing out of their individual value systems a value system for the "super-character" that represents them as a collective whole.

The theory aims to explain much about the role of *emotions* in human interactions. It also deals with the roles of *irrationality, deceit, disbelief, rational argument in the common interest,* and *morality.*

1.1 Formal preliminaries

The concept of a "frame" corresponds to that of the "game" of game theory, with the following differences. First, a frame is always subjective, representing the way in which the actors see their situation. Second, it is "soft"—i.e., capable of changing—whereas in game theory the game is taken as fixed.

The following formal definitions are taken over from part 1, where they are explained more fully.

A *frame* is an object

$$F = (Q, P) \tag{D1}$$

where the function

$$Q\colon X \to X \tag{D2}$$

is a *consequence* function from and to a set X of *outcomes*. Individual outcomes (members of X) are written x, y, z, etc. Qx shows, for each outcome x, the outcome that will actually be implemented if the characters attempt to implement x. The interpretation is that if $Qx \neq x$, then to form intention x is impossible; it becomes intention Qx.

When $QX \neq x$, x is called *infeasible*. (D3)

A feasible x is called a *future*. (D4)

The set of *outcomes* (the domain and codomain of Q) is the Cartesian product

$$X = \Pi(X_i \mid i \in C) \tag{D5}$$

of a family

$$(X_i \mid i \in C) \tag{D6}$$

of *strategy sets*. The index set C for this family is called the *cast*. (D7)

Individual cast members, written i, j, k, etc., are *characters*. (D8)

X_i is the strategy set of the character i. Members of X_i are written x_i, y_i, z_i, etc.

$$P = (P_i \mid i \in C) \tag{D9}$$

is a family of *preference relations,* one for each character i, defined over the set X of outcomes. Thus $P_i \subseteq X \times X$. "$(x, y) \in P_i$" means "i prefers x to y." It will also be written

$$x >_i y. \tag{D10}$$

Thus the negation of "$x >_i y$" is "i does not prefer x to y." It will be written "$x \leq_i y$" and taken to mean the same as "i potentially prefers y to x." Assumptions are:

1. $QQx = QX$ always; that is, every outcome has a feasible consequence.
2. $(x, y) \in P_i \Leftrightarrow (Qx, Qy) \in P_i$. Reason: preferences depend upon the futures Qx that outcomes x give rise to.
3. $(x, y) \in P_i \Rightarrow (y, x) \notin P_i$.

An *interaction* is an object

$$I = (p, f, x) \tag{D11}$$

where:
1. The family

$$p = (p^i \mid i \in C) \tag{D12}$$

is a family of *positions,* or publicly declared aspirations, for the cast of characters; that is, each p^i is a future which character i wishes to persuade others is its aspiration. We write

$$A = \{G \mid \exists i \in C\colon G = \{j | p^j = p^i\}\} \tag{D13}$$

for the corresponding partition of the cast into subsets sharing the same position.
2. The family

$$f = (f_G \mid G \in A) \tag{D14}$$

where A is the partition just defined is a family of *policies,* one for each group in A. The policy f_G is a function from X to X representing G's chosen pattern of reactions to the others' intentions; that is, if the characters in G perceive present intentions to be y, $f_G y$ is what G makes them by changing (or not changing) its intentions. This requires that

$$\forall y\colon f_G y \in [y_{-G}] \tag{D15}$$

where we draw attention to our use of the following notation:

$$-G = C - G \tag{D16}$$

(that is, if G is any nonempty subset of characters, i.e., what we shall call a *group,* then $-G$ is the complement of G in C);

$$y_G = (y_i \mid i \in G) \tag{D17}$$

(that is, if G is any group and y any outcome, y_G is the joint strategy that the group G has to implement in order to implement its part of y);

$$[y_G] = \{z \in X \mid z_G = y_G\} \tag{D18}$$

(that is, $[y_G]$ is the set of outcomes "offered" to $-G$ by G's choice of y_G. It is called an "offered set." Note in particular that $[y_C] = \{y\}$ and $[y\varnothing] = X$.)
The requirement D15 is, then, that G's reaction must be an outcome within the set offered by $-G$. Otherwise, it is not a reaction of G alone, but of others together with G!
3. x is a *confrontation point* or *fixed point,* meaning a particular outcome belonging to the intersection of all the characters' policies. That is, x is an outcome obeying

$$f_G x = x \text{ (all } G \in A). \tag{D19}$$

This fixed point x is also called a "threat point" or "conflict point."

Finally, the *informationally closed environment* of a dramatic situation is a set E of frames that represents all the different ways in which the characters might "frame" their situation without having any further information about it than they already collectively possess. (D20)

Building on these definitions, various theorems can be proved. We shall not set out the proofs, as they are trivial enough for the reader to supply, but **theorems will be indicated by the use of bold type, as here**. Though the proofs are trivial, it is important to point out which statements are theorems in order to show how the theory hangs together as a deductive system.

2. Dramatic resolution as agreement on a strict, strong equilibrium

Dramatic resolution was seen in part 1 of this article as going through five phases.

1. *Scene-setting:* The "author" creates a class E of possible frames from which the frame currently perceived by the characters is selected (see D20).

2. *Buildup:* A frame $F = (Q, P)$ is selected from E. Within this frame, each character i selects a position p^i (see D12). Thereby a partition A is defined, such that an element of A is a nonempty subset of characters (a "group") who all take the same position (see D13).
3. *Climax:* If all characters take the same position and it is a strict, strong equilibrium, phase 3 is skipped and the process moves to phase 4. Otherwise, each "group" G in A settles on a policy f_G (see D14), that is, on a pattern of reaction to the apparent intentions of the characters in $-G$. This leads all the groups to settle on a fixed point x that belongs to all their policies (see D19). At x, a confrontation takes place and paradoxes cause emotion. Characters may move to a new fixed point x or new policies f; by doing so they may change the nature, but cannot change the fact, of the paradoxes they face. Alternatively, they may change their positions, returning to phase 2 in order to do so. Alternatively again, their preferences P may change (see D9); they may reconceptualize their options X (see D5, D6); they may think again about the consequences Q that they see as following from different option combinations (see D2); or they may redefine the cast C of characters involved (see D7). These changes may lead to dissolution of paradoxes and hence to progress to stage 4. Otherwise, there is a return to stage 3.
4. *Resolution:* All characters having taken as their position the same strict, strong equilibrium, this is adopted as an understanding between them. In exploring its details, they may uncover further problems. Otherwise, they move to the dénouement.
5. *Dénouement:* The understanding between the characters is implemented, possibly leading to new dramatic confrontations.

We begin our more formal discussion of this process by defining the game-theoretic solution concept of "strict, strong equilibrium." This is what a "dramatic resolution" is required to be.

2.1 Definition of strict, strong equilibrium

Drama theory hinges on the fact that if and only if no paradoxes of rationality exist, then the characters have resolved their problem in a totally convincing manner. Specifically, in phase 2, when the characters decide on a family of positions, we shall show that this confronts them with paradoxes if and only if it does not *determine a unique position that is a strict, strong equilibrium.* What does this mean?

Let us introduce some general concepts.

Improvements. Define the set of *potential improvements for* a group H *from* an outcome x as the set:

$$M_H x = \{y \in [x_{-H}] - Q^{-1}Qx \mid y \geq_H x\} \qquad \text{(D21)}$$

and the set of all potential improvements from x, for any group H, as the set

$$Mx = \cup(M_H x \mid H \subseteq C). \tag{D22}$$

A potential improvement from x is, as the name implies, an outcome y that the members of H can move to by changing their own intentions and that is potentially better for them all than x—i.e., the slightest increase in preference for y would cause them all to benefit from the move.

A potential improvement y is *strict* if $y >_H x$. **A strict improvement is such that all characters in H definitely benefit from it.** (D23)

Though all characters in H benefit from a strict improvement, the complementary set $-H$ might, if they notice H's intention to carry out the improvement, react in such a way that for some member of H the benefit disappears. To capture this, define a *sanction* against the potential improvement y for H from x as an outcome z such that

$$z \in [y_H];\ \exists i \in H\colon z \leq_i x \tag{D24}$$

and define a *guaranteed improvement* as a potential improvement against which there is no sanction. (D25)

We have then that **a non-strict improvement is a sanction against itself**, and consequently **a guaranteed improvement is strict.** Also, **y is a guaranteed improvement for H from x iff**

$$[y_H] >_H x$$

where the convention used is that "$Y >_H x$," where Y is a *set,* means that every outcome in Y is preferred to x to by the members of H.

Clearly, **all characters in H benefit from a guaranteed improvement, regardless of any further reactions by $-H$.**

Strict, strong equilibria; strong equilibria; the core. A *strict, strong equilibrium* is now defined as an outcome x such that

$$\mathrm{Mx} = \varnothing \tag{D26}$$

—i.e., there are no potential improvements from it. **This means that any group able to move from a strict, strong equilibrium to a nonequivalent outcome contains at least one member who would lose by the move!** A strict, strong equilibrium is thus very stable in that, if characters expect each other to implement it, their expectations will reinforce each other. Hence, a strict, strong equilibrium is "honesty-reinforcing"; that is, no subset of characters will want to deceive others into believing it intends a strict, strong equilibrium. If it wants them to believe this, it is because it is so.

Not necessarily so stable is a *strong equilibrium,* defined as an outcome from which there are no strict improvements. (D27)

Least stable may be an outcome that merely belongs to the *core*, defined as the set of outcomes from which there are no guaranteed improvements. (D28)

Clearly, **the strict, strong equilibria are strong equilibria, and the strong equilibria are members of the core.**

The core $(= \{x \mid \sim \exists\, y_G\colon [y_G] >_G x\})$ has become increasingly important in game theory. Despite possible lack of stability, it has been favored as a solution concept, because it corresponds to the concept of competitive equilibrium in economics. It can be empty, but unlike the sets of strong and strict, strong equilibria, it is nonempty in many important situations.

2.2 *Dramatic resolution*

Let us now return to a discussion of phase 2, the buildup phase, where the characters choose a family

$$p = (p_i \mid i \in C)$$

of positions.

Clearly, if all these positions coincide at a strict, strong equilibrium, the characters have resolved the issues among them and can go on to phase 4. If they merely coincide at a strong equilibrium, there is a worry that a group might form and be encouraged—the slightest change of preferences would be enough—to find an improvement. The position is less stable.

If merely a member of the core, the position is stable to the extent that each group G with an improvement y can be discouraged from making it by the threat of a sanction wielded by the other characters, since **x belongs to the core iff**

$$\forall\, G\colon \forall y \in M_G x\colon \exists\, z \in [y_G]\colon \exists\, i \in G\colon x \geq_i z$$

—in words, iff every improvement is deterrable by a sanction.

Dramatic resolution of the problem by all proposing a strict, strong equilibrium does, however, depend on all proposing the *same* strict, strong equilibrium. A frame may, as in the "chicken" game of figure 1, have more than one strict, strong equilibrium; if two characters propose different ones, they have a problem still.

We will look at this problem in section 4. In the next section we look at the paradoxes that may exist when characters do agree on a single position, but that position is not a strict, strong equilibrium.

3. Paradoxes of cooperation and how they are overcome

Our first paradox is exemplified by the prisoner's dilemma game. We will discuss it at length. Much of the discussion, since it concerns the general way in which paradoxes may be overcome, will apply to later paradoxes also.

Suppose, then, that the characters in phase 2 have taken up a family p of positions. We have already (D13) defined the set $A = \{G \mid \exists\, i\colon G = \{j \mid p^j = p^i\}\}$, **which**, from this definition, **is a partition of *C* such that the members of each group $G \in A$ share a unique position p^G distinct from the position p^H of any other group *H*.**

The family $(p^i \mid i \in C)$ of individual positions thus determines a family $(p^G \mid G \in A)$ of group positions. If there is disagreement between the positions taken, this family has more than one member. If all agree, it has as its sole member the single position taken by the whole cast *C*.

Whether there is total agreement or not, there can now exist for a group in A the *cooperation paradox,* illustrated by the "prisoner's dilemma" game at the top left of figure 1. In this game, as is well known, the definition of "rational choice" is straightforward; it is rational to choose strategy 2, regardless of the other's choice. The paradox is that *if both players are rational* (thus obtaining (2, 2)) *both are worse off than if both are irrational* (and obtain (3, 3)).

Confronted with this situation in a drama, we would expect both players to take (3, 3) as their position. We would then have $A = \{C\} = \{\{\text{row chooser, column chooser}\}\}$ and $p^C = (\text{row 1, column 1})$. Each player would tell the other of its intention to choose strategy 1 and its expectation that the other will do so. The problem is, how can intention or expectation be genuine, given that it is irrational?

Credibility. The paradox shows that members of a group $G \in A$ may have a serious problem in assessing each other's *credibility* and convincing each other of their own. They may be taking as their position a future it would be "irrational" to intend. How can they believe each other if success in achieving their position would give reason to believe that some did not genuinely intend to carry it out?

The general credibility problem illustrated by prisoner's dilemma is that *members of a group may have potential improvements from the group position.* Clearly, **this cannot happen if the group position is a strict, strong equilibrium**.

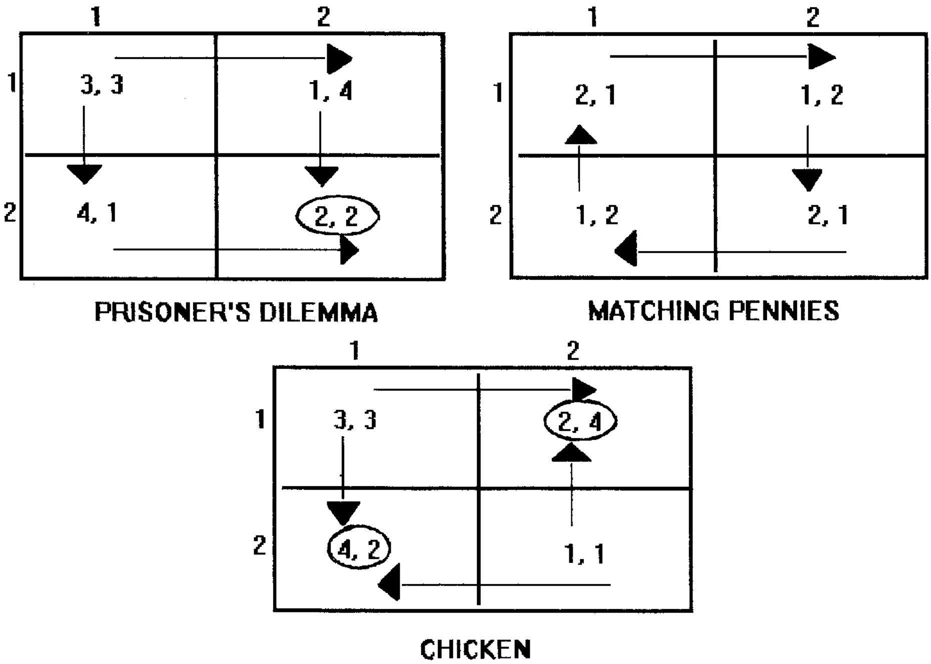
It can quite well happen in general, however. And this is the paradox of cooperation—the fact that a member of a group may quite reasonably take up a position from which it has a potential improvement, and is therefore tempted to defect.

In prisoner's dilemma itself, individuals have improvements in moving from (3, 3), and from the off-diagonal cells to (2, 2). There is also an improvement for the group C in moving from (2, 2) to (3, 3). There are in fact strict improvements from every cell!

Temptation. The paradox of cooperation may be put in another way. Let p^G be a group position within the family $(p^G \mid G \in A)$ of positions. If

$$x \in Mp^G, \text{ and } x \geq_i p^G \qquad \text{(D29)}$$

for some $i \in G$, call the pair (x, p^G) a *temptation* for *i* relative to p^G. Then the paradox may be stated by saying that temptations may exist for one or more members of *G*.



PRISONER'S DILEMMA

	1	2
1	3, 3	1, 4
2	4, 1	2, 2

MATCHING PENNIES

	1	2
1	2, 1	1, 2
2	1, 2	2, 1

CHICKEN

	1	2
1	3, 3	2, 4
2	4, 2	1, 1

Figure 1. Three paradoxical games. One of the players chooses row 1 or 2 and the other column 1 or 2. Payoffs for row chooser and column chooser respectively appear in each cell. Arrows show individually rational moves (i.e., changes of intention); circles show cells where both players are individually rational. In these games each outcome is a future, that is, $Qx = x$ for all x.

This creates a credibility problem. How, if they succeed in getting acceptance of their position, can they trust each other to resist temptation?

In the prisoner's dilemma example, given the virtually inevitable joint position at (3, 3), (4, 1) is a temptation for row chooser and (1, 4) a temptation for column chooser.

Preference change, preference friction, and emotion. The dramatic hypothesis is that when a group confronts this paradox, a member who suffers from a temptation will feel pressure to *change its preferences* so as to eliminate the temptation. Pressure will come both from other members of the group—who will try to persuade or induce the character to renounce the temptation—and from itself, as it tries to make others believe that it really would implement p^G. But how can a character change its preferences? Not by simply deciding to do so!

Preference friction we define as the set of factors that make it difficult for a character to change the real-world costs and benefits, and the system of values by which they are weighed, that underlie its preferences for one future over another. (D30)

Because of preference friction, a change of preferences requires *emotional energy,* which may be positive toward other people (love, sympathy, goodwill, etc.)

or negative (anger, rejection, contempt, hate, etc.). In the case in which a promise is being made to adhere to a joint position, the required emotion is positive toward others in the group (showing love, sympathy, good will, solidarity, etc.) and negative toward non-members of the group who prefer the improvement (expressing rejection of their values, lack of sympathy, contempt, etc.). Expressing the appropriate emotion has the function of enhancing the credibility of a character's preference change as well as making it possible.

The role of emotion in solving the cooperation paradox can be summed up in this statement: *positive emotions toward the recipient of a promise have the function of making the promise credible; so do negative emotions toward those who would like the promise broken.*

Irrationality. While preferences are changing, behavior is without a proper foundation in a system of values and preferences, and is in that sense irrational. We can clarify the function of irrationality if we define the transformation $P_i \rightarrow P'_i$ that is taking place; irrational behavior is simply behavior in accordance with the preferences P'_i that are attempting to take over from P_i, rather than in accordance with existing, established preferences P_i. Irrationality is thus a dynamic phenomenon accompanying preference change and emotion.

If preference friction is so great as to preclude actual change from happening (as when a person wishes he or she were dead in order to help or hurt another, but does not seriously contemplate suicide) then behavior may be irrational in that it is temporarily in accordance with P'_i; when emotion subsides, P_i will reassert itself. If in such a case the carrying out of the threat or promise depends on short-term rather than long-term preferences (as it does, for example, when a gun is pointed), then irrationality is actually a substitute for preference change; on its own it can make the threat or promise credible, since it can make the recipient think "They're crazy enough to do it!"

Preference reversal. Can we specify the transformation $P_i \rightarrow P'_i$? To some extent we can.

The immediate pressure, given a temptation (x, p^G), is to transform P_i (defined in D9 as a set of preference pairs) by deleting from it the preference pair (x, p^G)—if it is in P_i—as well as any pair equivalent to it, and adding the pair (p^G, x) and its equivalents.

Call this transformation of P_i the "reversal" of (x, p^G). After it, whether or not i previously preferred the improvement to the position, it now definitely prefers the position to the improvement!

For a general definition, take any pair $(y, x) \in X \times X$ and define the *reversal* Rev(y, x) as the following transformation of a preference relation P_i.

$$\begin{aligned} P'_i &= \text{Rev}(y, x).P_i \\ &= (P_i - Q^{-1}Qy \times Q^{-1}Qx) \cup Q^{-1}Qx \times Q^{-1}Qy. \qquad \text{(D31)} \end{aligned}$$

We have then that **if P_i is asymmetric, so is Rev$(y, x).P_i$.**

Rationalization. Reversal relieves the immediate pressure for preference change. But, though it preserves asymmetry, it may transform P_i from an ordinal into a non-ordinal relation, that is, one with intransitivities. These are themselves a kind of irrationality!

There is a deeper sense in which simple reversal is irrational: it is arbitrary. It does not flow from a proper evaluation of the costs and benefits of alternative futures.

Because it is irrational, the dramatic hypothesis is that mere reversal is sustained by emotion and lasts only as long as the emotion lasts. When emotion subsides, the old preferences P_i reassert themselves. To become permanent, reversal needs to be followed by another transformation which we call *rationalization*. Unlike reversal, this depends not upon the abstract preference relation P_i but upon the real-world costs and benefits that underlie it.

Reversal is an abstract requirement that a character must try to make convincing, to itself and others, through a rational reconstruction of its value system, as when a character decides that it loves another, and can therefore rationally change its preferences to give higher value to futures which benefit the other. This rational reconstruction is rationalization.

Rationalization cannot be given a general formal definition as it depends upon the concrete facts of the situation, i.e., upon the characters' underlying reasons for their preferences and the arguments by which they seek to change them. It may, by changing the whole system of weights used in evaluating futures, lead to changes going beyond the immediate requirement to eliminate a temptation, as, for example, when a character adopts a lifelong code of loyalty to an organization in order to obtain concrete near-term benefits from it. (D32)

Deceit and disbelief; the paradox of belief. The credibility problem we are discussing consists in one character suspecting another's assertion that it has or would have a certain intention. Why does it suspect? Because *deceit* would benefit the other more than truth telling. Hence *disbelief* is indicated.

Why then should anyone in a paradoxical situation ever believe anyone? This is the paradox of belief. If A tells B something, it is because it is A's interest that B believe it. That, however, in a paradoxical situation, gives B reason not to believe it! The mere fact of communication—as distinct from the content—is in such situations a reason for disbelief.

Communication accompanied by emotion may possibly alleviate this problem. Emotion can be simulated, but not, apparently, very well (see Frank 1988, chapters 6–7, for a review of the evidence). Thus emotion is a partial solution to the paradox of belief. Observing others' emotions is a better-than-chance way of establishing their credibility.

Observation of emotion should here include observing the absence of emotion, which conveys the message that the non-emoter is not faced with a rationality paradox, and can therefore be trusted. "I have no improvements from p^G" is the claim implicitly made by a member of G who shows no emotion when making the promise to adhere to the group position p^G.

Though emotion or its absence may be hard to simulate—particularly for organizations, whose internal workings may be visible—it is certainly not impossible. Moreover, strong emotion can carry an ambiguous message, since it indicates that the emoter is striving to overcome preference friction without indicating that it has succeeded. "The lady doth protest, too much, methinks," says a Shakespearean character.

The paradox of belief, moreover, undercuts all these devices. It means that all the methods of trying to establish credibility so far discussed may have the opposite effect and cause disbelief, since the subtext of each is that the sender of the message has a reason to induce the receiver to believe it.

Rational arguments in the common interest. The paradox of belief is solved by paying attention to the real-world content of a communication, rather than to the mere fact that it is made or to its abstract purpose (to persuade others to resist temptation and make them believe one will do so oneself).

The content can appeal not to emotion or its absence but to *evidence* and *rational argument*. To be effective, these should be presented with appropriate emotion.

Use of evidence and rational argument does not imply lack of values; on the contrary, it assumes the existence of certain fundamental values common to sender and receiver. Where are these to be found? There is a general answer to this. The fact that group members have a common position means that they necessarily have common values. These common values may however be waiting to be invented or defined by looking at and generalizing about the concrete common interests that underlie their choice of the same position.

The supercharacter; morality. Construction and elaboration of common values in order to persuade members to believe in each other, be worthy of belief, and pursue a common end constitutes the building up of a *super-character*—a higher level character composed of the group members working together. It is also, if accompanied by genuine and appropriate emotion, the practice of *morality* in concrete circumstances.

Thus drama theory presents itself as a theory of practical ethics as well as a positive theory. It is both because it is a theory of how people naturally resolve their differences.

Is such morality relative? Yes. But though it may be the morality of thieves, monopolists, or other subgroups working against the larger interest, relative morality is moral at its own level, that is, within its own frame. Its relativity is alleviated to the extent that, in building up a common interest within the group, group members may appeal to the interests of the larger community to which all belong. The values of this larger community are an obvious source of common values for the members of the group to draw upon.

Implementation in stages. A factor influencing the amount of preference friction to be overcome in making a promise credible is whether the characters' options will eventually be implemented simultaneously or in stages (recall that in defining a frame, in part 1 of this article, we left it open whether the *dénouement* would occur in stages, and, if so, what these would be). If in prisoner's dilemma, for example, row chooser's choice is implemented after column chooser's has become irreversible, then after column 1 has been implemented, row chooser will be left with a one-person decision—at a time when the pressure for preference change that existing during pre-play negotiations will have disappeared. Foreseeing this, both players know before anything is done that unless row chooser's preference change in favor of (3, 3) is thoroughly rationalized and assimilated, it will—assuming the first column is chosen—be reversed before implementation. There is great preference friction to overcome in making or believing in such a change.

On the other hand, if a character's choice becomes irreversible before or at the same time as others', the pressure for preference change in regard to that choice continues until the point at which it becomes irreversible. There is less friction to overcome in making or believing in such a change!

Changes in options or consequences. What happens when characters have to recognize that preference friction is too great to allow them to make, or to make credible, the required preference changes? They may creatively think up other kinds of change. They may find a way of eliminating options, or assigning consequences to them, so that ineluctably preferred improvements are wiped out: in prisoner's dilemma, for example, if row chooser cannot prefer the (3, 3) cell to (4, 1), a way may be found to make row 2 disappear, or a consequence may be found linking (4, 1) to (2, 2)! Odysseus had himself tied to the mast to eliminate his *option* of following the Sirens; he did so because he knew that their music would cause him to forget the *consequence* of doing so.

Alternatively, new options may be added enabling acceptable replacements for the current position to be reached by strategies which do not offer destabilizing improvements. In prisoner's dilemma, column chooser might see how to add a column containing a cell leading to much the same future as the present (3, 3) position without exposing row chooser to temptation. In adding new options, characters may even add to the cast C—as when a business whose shares are being bought calls in a "white knight."

Changes in position. As just noted, when all else fails, characters may substantively change their positions. This may mean "giving up" or trying to achieve the same objectives in a different way. In prisoner's dilemma, both might decide that they can only hope for (2, 2).

Of course, in order to accept this as a dramatic resolution, they now have to give up the temptation to move jointly from (2, 2) to (3, 3)! Otherwise, they will feel that the resolution is not fully satisfactory. The drama will have ended on a false, unsatisfying note.

This unsatisfactoriness is explained by the fact that (2, 2) is not a strict, strong equilibrium, since there is a joint improvement—a temptation—for the two characters to move to (3, 3).

Another kind of cooperation paradox. The "cooperation paradox" that we have discussed so far consists in the fact that a character may have a potential improvement from its position. We now need to point out another kind of paradox of cooperation that may exist, with or without the first. It consists in the fact that when more than one position is taken, a character may potentially prefer another position to its own. This could happen in prisoner's dilemma if one character, but not both, despaired of making (3, 3) mutually credible, and hence adopted (2, 2) as its position.

To put this more generally: a character would normally be expected to prefer its own position to any other. But if, after it has abandoned its position, others in its group refuse to follow, it may find that this is no longer the case. In order to get round one breakdown of rationality, it has to commit itself to work against, and to persuade and pressure others to renounce, a future that it may prefer in favor of one that it does not prefer!

Is this not irrational? We classify it as a second kind of cooperation paradox. Just as the existence of an improvement casts doubt on the genuineness of a character's intention to implement its position *once agreement is reached,* so a preference for another's position casts doubt on its determination to stick to its own position *while agreement is being sought.*

In each case, the problem lies in the existence of a "temptation" for an individual i $\in$ G: in the first case the temptation is a preference (y, p^G), where $y \in Mp^G$; in the second case, a preference (p^H, p^G).

The difference in timing is this: if *i* succumbs to temptation, it would succumb to the first kind of temptation at the *dénouement* phase (phase 5), by failing to *implement* p^G; it would succumb to the second kind at the climax phase (phase 3), by failing to *fight for* p^G.

Accordingly, we will call the first a *phase 5* temptation, the second a *phase 3* temptation. (D33)

The phase 3 cooperation paradox is solved by a character changing its preference so that it *does* prefer its position to all others. First it applies to P_i the transformations

$$\text{Rev}(p^j, p^i), \qquad \text{for } p^j \neq p^i.$$

These are followed if necessary by rationalization. Rationalization, as before, provides justification for the new preferences on the basis of a coherent system of values and, at the same time, gets rid of any intransitivities.

The emotions, rational arguments, etc., that accompany these transformations will, we hypothesize, be as we have described in the case of the phase 5 cooperation paradox—so will the resort to reframing options, consequences, or cast. Note, however, that when the temptation is only a phase 3 temptation, not a phase 5 one, then implementation in stages cannot create any preference friction.

Often the two kinds of cooperation paradox appear together: one and the same temptation exemplifies both, as we have seen in the prisoner's dilemma case when one character, but not both, gives up trying for (3, 3) and takes up (2, 2) as its position. The two problems can occur separately, however. Take chicken (third game in figure 1). Suppose that row chooser takes (4, 2) as its position and column chooser takes (3, 3). Suppose that row chooser gives in and changes its position to (2, 4), skipping over (3, 3) on the grounds that it cannot trust column chooser—or, perhaps, itself—not to defect from there. It then faces a phase 3 paradox but no phase 5 paradox, since it has no improvement from (2, 4).

Consistency and unity. If a member of G is tempted by an improvement from its position, or if it prefers another's position to its own, its preferences are inconsistent with its position. It suffers therefore from "cognitive dissonance" (Festinger 1957).

Accordingly, the position p^G within the family $(p^G \mid G \in A)$ will be called *consistent* when it is subject to no such temptation or preference—i.e., just when

$$p^G >_G Mp^G \cup \{p^H \mid H \in A - \{G\}\}. \tag{D34}$$

A family $(p^G \mid G \in A)$—and the underlying family $p = (p^i \mid i \in C)$, which contains the same positions—will be called *consistent* when all its members are.

A family of positions suffers, obviously, from another kind of dissonance if it contains multiple positions. Call the family $p = 0$ $(p^i \mid i \in C)$ *united* when its elements p^i are all equal to one another—**implying that it determines a single position** p^C—and *multiple* otherwise. (D35)

First theorem of the final state. The dramatic hypothesis is that inconsistent or multiple positions generate emotional pressures for transformation of preferences, options, cast, consequences, or the positions themselves. If successful, these pressures lead eventually to a consistent, united family of positions.

We now point out that **a consistent, united family of positions contains a single, strict, strong equilibrium**. Consequently, a consistent, united family will go straight from phase 2 to phase 4, the phase of dramatic resolution.

This "theorem of the final state" shows the state of characters' expectations when a dramatic resolution is finally obtained. It shows also that we have answered part of the question, namely, how is this final state reached? Characters who share a single position reach dramatic resolution by making their position consistent with their preferences.

4. How the paradoxes generated by multiple positions are overcome

We must now address the question: what do characters do when they do not share a single position—when they are disunited? In this case they have to face and overcome various other paradoxes.

In order to resolve their problem, the dramatic hypothesis is that they must get into an "interaction" as defined in D11. This means that each group G sharing a common position follows a "policy" f_G as in D14, with the result that the groups confront each other at a fixed point x as in D19. But for this to happen, it must be the case that the policies f_G do in fact intersect at some point or points x.

Nonintersecting policies confront the characters with a paradox we call the "paradox of indeterminacy."

4.1 The indeterminacy paradox

To see how this paradox arises, suppose we have a multiple family $(p^G \mid G \in A)$ of consistent positions. The groups try to interact by announcing intentions to each other and attempting to guess whether each other's intentions are genuine. They then react—or fail to react—to each other's perceived intentions.

Assuming that each begins sooner or later to follow a consistent plan, we model this by supposing that the reactions of each group G are in accordance with a specific function f_G: $X \to X$ that yields, for any outcome x, the outcome $f_G x \in [x_{-G}]$, which G will move to if it believes that x is cointended. But how are these reaction patterns implemented? Do the groups react in a fixed order to a given contention x, or all at once? We cannot say, in general. All we can say is that the situation cannot settle down to allow characters to communicate and assess each other's credibility unless it reaches an outcome belonging to all f_G, that is, an x, the "confrontation" or "fixed" point of the interaction, which belongs to all f_G—i.e., an x obeying D19.

We define a family of policies as *determinate* just when they have such a fixed point.

The *indeterminacy paradox,* illustrated by the game of matching pennies—second in figure 1—is that for a given set of policies no such point may exist! That is, we may have a family $(f_G \mid G \in A)$ of policies such that

$$\sim\exists x \colon \forall G \colon : f_G x = x. \qquad \text{(D36)}$$

In matching pennies, for example, row chooser's obvious policy will be to choose the same as column, while column chooser's policy will be to choose differently from row. Both cannot succeed!

Reification. The indeterminacy paradox is more technical than the others discussed in this article. Faced with it, we hypothesize that groups become frustrated at their inability to carry out their policy: each time they have momentary success, they find or suspect that others' intentions have changed and they must try once more.

One commonly adopted solution is in some way *to "reify" the indeterminate outcome.* In matching pennies, for example, a player might decide to resolve the indeterminacy by choosing heads or tails at random, with a 50 percent probability of choosing each. The indeterminate outcome is thereby "reified." It becomes a well-defined thing: a lottery that the player has a 50 percent chance of winning. This stochastic method of reification is the basis of von Neumann's concept of mixed strategies. In matching pennies it amounts to each player thinking up a new strategy to add to its existing two—the strategy of tossing a coin to decide on a choice of row or column.

An alternative method is to *redefine the policies themselves—in whole or in part—as strategies*. In matching pennies, row chooser might consider itself and column choosee as having the additional strategies "try to guess what the other will choose and match it" and "try to guess what the other will choose and not match it." Row and column chooser thus consider themselves to have three strategies each, not two; they assign a future (perhaps stochastic) to the outcome in which their two extra strategies are chosen, and assign preferences to this future. This may be regarded as another form of *reification.* (D37)

By thus formally extending their strategy sets, groups redefine the situation so that it can be regarded as determinate. After this redefinition it is, in fact, determinate, since in the new frame in which these extra strategies (stochastic or functional) appear, there does exist an x obeying D19.

4.2 Inducing policies and realistic positions

The remaining paradoxes that characters may face are concerned with the general problem of "inducement." We begin by discussing this concept.

Inducement. Let us ask: Why do groups pursue policies? Why react? Why not allow intentions to settle by keeping to the same intention when other groups' perceived intentions change? It is simple: if a group $G \neq C$ allows intentions to settle at an outcome preferred by all in $-G$ to p^G, it effectively gives up its position. It cannot allow this; it has to move away from such an outcome, otherwise *de facto* it has changed its position, and should go back to phase 2!

Formally, we are saying that if A contains more than one group, then each group must follow a policy that "induces" its position p^G, where a policy f_G is said to *induce* a future x if

$$\sim\exists y: f_G y >_{-G} x. \qquad \text{(D38)}$$

(Note that according to this definition, the group C consisting of the whole cast cannot have an "inducing" policy.)

Why the term "inducement"? Because by pursuing an inducing policy, G necessarily reacts to any cointention y in a way that potentially pressures at least one character in $-G$ to abandon its position, adopt the position p^G, and join the group G. This is so, since **f_G induces p^G iff**

$$\forall y: \exists i \in -G: p^G \geq_i f_G y.$$

We conclude that merely having a position in a sense commits a proper group G (i.e., a proper subgroup of C) to a campaign of trying to convert members of the complementary group $-G$ to change their positions and adopt G's. This campaign is conducted through an inducing policy f_G.

We have the theorems:

G ($\subset C$) can have a policy for inducing x iff $-G$ has no guaranteed improvement from x.

x belongs to the core iff it is Pareto optimal (no future exists that is better for every character) and every group $G \subset C$ has a policy that induces it.

Realistic positions. Consider now a group G that is trying to induce its position. The following question arises if the complementary group $-G$ contains two or more characters: when G has converted at least one character to its cause, will it (the enlarged group) have a policy with which to induce at least one further character (from the reduced complementary group) to adopt its position, and so on till all are converted?

The group's position may be called "realistic" when the answer is yes. Accordingly, we define a future x as *realistic* for a group G when no group contained in $-$G has a guaranteed improvement from it. (D39)

We then have the theorems:

A future is realistic for G iff it is realistic for every group that contains G, and iff every such group has a policy for inducing it.

A future belongs to the core if it is realistic for every one-member group and only if it is realistic for every group.

Every future is realistic for the universal group C.

A future x is realistic for G iff G contains a representative from every group that has a guaranteed improvement from x.

4.3 The deterrence paradox

The definition of a "realistic" future enables us to present our next paradox. It is summed up by the Latin motto *si vis pacem para bellum*: if you wish for peace, prepare for war. This is the paradox that confronts a group with a non-realistic position. It is in a situation where, in order to obtain one future, it must think up and be ready to implement another!

This may not, perhaps, seem paradoxical to game theorists. To many it seems unbearably so, as countless arguments between pacifists and realist politicians attest. The difficulty in modeling it game-theoretically is that it concerns whether or not to exclude certain strategies—e.g., those that lead to war—from consideration. How? A game is fixed! Strategies that exist can't be excluded!

In drama they can. Suppose we start with the "chicken" frame in figure 1. Let row chooser's initial position be (3, 3); column chooser's (2, 4). Row chooser could make its position consistent by reversing its preferences, so that it would prefer "(3, 3)" to "(4, 2)."

This may, however, be difficult to achieve or make credible due to preference friction. But an alternative may exist. Suppose that the environment E contains, in addition to the whole chicken frame, the same frame with row 2 deleted. This means that instead of changing its preferences, row chooser could make its position consistent by "disarming," i.e., by ceasing to be ready or able to implement row 2. This might seem a better way to achieve consistency, as it takes away the opportunity to defect. The result, however, is that the position (3, 3) becomes unrealistic, since (2, 4) becomes a guaranteed improvement from (3, 3) for column chooser!

The paradox is that row chooser's position is realistic only by virtue of the existence of a strategy it does not aspire to use.

What is the general point we are making? Stated generally, a *deterrence paradox* may be said to confront a group G that takes a position p^G that, within its current frame, is unrealistic:

$$\exists\, y_H\colon (\varnothing \subset H \subseteq -G)\colon [y_H] >_H p^G. \qquad \text{(D40)}$$

Such a group, if it tries to maintain its position, will find itself under emotional pressure to find a "deterrent" strategy that makes p^G realistic. Let us explore this.

Emotion, option change, deceit, disbelief, rational argument, and morality; the phenomenon of "demonization." A group that takes an unrealistic position may, of course, change its position. If it does not, they must think up—create—a "punishment" for H in the form of a new strategy y'_{-H} such that, in the transformed frame F', the tuple (y_H, y'_{-H}) is not preferred to p^G by at least one member of H.

The emotional energy needed for this creativity is, we hypothesize, provided by negative feelings toward members of H—hostility roused by fear of their putative action. Once a deterrent strategy y'_{-H} has been "sketched in" sufficiently to fix the members of H (say H') that are to be punished by it, these negative feelings are concentrated on them. "Demonization" of H'—i.e., attribution of bad characteristics to them as the potential enemy for whom y'_H has to be prepared—is the appropriate rationalization. (D41)

An example to illustrate this was given in part 1 (Howard 1994) where we discussed a simplified model of the conflict in Bosnia. In that model, the Western alliance had adopted an unrealistic position in demanding that Serbia leave Bosnia. Accordingly, a process of demonization of Serbia was taking place in Western countries as they tried to "psych" themselves up to consider the option of armed intervention.

We see, then, the role of emotion in bolstering option change, specifically, in bolstering the introduction into the frame of a deterrent option. What, though, are the roles of deceit, disbelief, rational argument, and morality? We hypothesize that at this stage, the stage of modifying the frame so as to introduce or create a deterrent option, rational arguments and morality are used *internally* to combat *internal* manifestations of the paradox of belief. That is, they are used to combat deceit and disbelief affecting communication channels between characters in the group *G* and within those characters themselves. (Recall that we make the assumption that each character is a "supercharacter" made up of internal mini-characters, among whom communications have to take place.)

In sum, the deterrence paradox is resolved either by characters adopting new positions which are realistic within the given frame or by their extending the frame to incorporate deterrent options making their given positions realistic.

Resolution of this paradox affects the core. We have said that the core of a frame may be empty. If however any group has adopted a position that is both consistent and realistic—reframing the situation as necessary in order to do so—then we can be assured that the core is nonempty. We have the theorem **a consistent, realistic position belongs to the core**.

4.4 An example: "split the dollar"

Before going further, let us illustrate what has just been said by applying it to a frame with an empty core. We can then see how the dramatic process might make the core nonempty.

We will take the well-known game of "split the dollar" and treat it as a drama. Three characters, A, B, and C, can have a dollar (imagined to be continuously divisible) if a majority of them can agree how to split it between them. Whatever split is agreed, there is a guaranteed improvement from it for some group of two: thus (1/3, 1/3, 1/3) can be improved upon by A and B going to (1/2, 1/2, 0), which can be improved upon by B and C going to (0, 2/3, 1/3), and so on.

Most of the action in this game consists of taking and retaking positions. Suppose the position of all is (1/3, 1/3, 1/3). A and B may be unable to resist the temptation to intend (1/2, 1/2, 0) instead; but if this is their position, C can try to tempt B to defect to (0, 2/3, 1/3).

At this point, as we are in a drama rather than a game, we may suppose that B realizes that this offer too would be subject to temptation, and demands more of C than a temporary change of intentions. If intentions are to fix at (0, 2/3, 1/3), says B, temptations to defect must be convincingly eliminated. With the possibility of deceit on everyone's mind, emotional signs are now scrutinized to assess credibility, and rational arguments appealing to moral criteria such as fairness assume importance.

If, finally, these tendencies toward dramatic resolution succeed, a particular position, such as (1/2, 1/2, 0), is made consistent by changes in the preferences of A and B, causing them to prefer it to any temptation. But how can they make such changes? Mere reversal creates intransitivities through the fact that A (or B) still prefers a redistribution between A and B themselves that would give it more, as in:

$$(1/2, 1/2, 0) >'_{\mathrm{B}} (1/2 - \mathrm{x} - \mathrm{y}, 1/2 + \mathrm{x}, \mathrm{y})$$
$$\geq_{\mathrm{B}} (1/2 - \mathrm{x}, 1/2 + \mathrm{x}, 0) >_{\mathrm{B}} (1/2, 1/2, 0).$$

The change needs to be rationalized to get rid of such intransitivities. The obvious solution is to make (1/2, 1/2, 0) preferred not only to temptations but to redistributions between A and B; this makes it a realistic and consistent position while restoring ordinality to A's and B's preferences, and it can be morally justified as being "fair" between A and B.

What position can C now realistically take? Only the same! **(1/2, 1/2, 0) is now the only realistic position for C also!** C thus encounters the deterrence paradox. Angrily it looks for means of revenge—new strategies by which it can punish A, B, or both. Should it find none, it must reconcile itself to the inevitable and adopt the position (1/2, 1/2, 0) or nothing.

This is already a strict, strong equilibrium. With C reconciled to it (C **does not need to change preferences, just agree**), it becomes the dramatic resolution of this particular play of split the dollar. (Note that there is no requirement in drama theory that the same resolution be reached every time.)

We submit that this is a more convincing account of how people play split the dollar than that offered by game theory.

4.5 Confrontations and mutual inducement

Suppose now that our characters, having taken a family $p = (p^i \mid i \in C)$ of individual positions determining a family $(p^G \mid G \in A)$ of group positions, have, if necessary, solved the paradoxes of indeterminacy and deterrence. This has enabled them to form a determinate family $f = (f_G \mid G \in A)$ of group policies that induce their positions. Where has all this got them?

It has enabled them to stage a confrontation! Being determinate, the policies contain at least one fixed point x, so that and after some initial "to-ing" and "fro-ing" the characters can settle at such a point. We now have what we define as a *confrontation,* that is, an object

$$(Q, P, p, f, x), \tag{D42}$$

where:

(Q, P) is a frame;
(p, f, x) is an interaction taking place within (Q, P);
the policies f_G induce the positions p^G;
the policies are determinate with a fixed point at x.

Note that it is not part of the definition of a confrontation that the positions be consistent or multiple. As a degenerate special case, we can have a *united* confrontation. Here all characters share a single position p^C—as they do in prisoner's dilemma. The policy f_C is in this case more or less redundant: **it may be any function from X to X that has a fixed point at x.** If the confrontation is not only united but consistent, we know (it is our first theorem of the final state) that everyone is proposing the same strict, strong equilibrium.

In a multiple, consistent confrontation, on the other hand, **the fixed point x obeys**

$$\forall G: \exists i \in -G: p^i >_i p^G \geq_i x.$$

Consequently, in a multiple, consistent confrontation each group G is implicitly trying to persuade at least one non-member i to accept the group position p^G on the grounds that it is just as good for it as x—though admittedly worse than p^i!

Define the set of those being thus persauded—i.e., the set

$$\{i \mid p^i >_i p^G \geq_i x\}, \qquad \text{(D43)}$$

(which is nonempty if the confrontation is multiple and consistent) as the group *being induced* by G, or G's *induced group.* In a multiple, consistent confrontation with two characters, this induced group merely consists of the other character; in the three-or-more person case, it need not be a group in A—i.e., it need not be a group sharing a common position. Its members may be drawn from various groups.

In the case in which an individual i in the induced group is indifferent between p^G and x, the slightest encouragement—meaning, the slightest credible indication of an extra benefit—is sufficient to make p^G strictly preferred to x. We hypothesize therefore that G will try to give i such credible encouragement—provided, of course, that x is not equivalent to p^G. And if it is? **If x is equivalent to p^G in a multiple, consistent confrontation, then there are more than two groups, no member of G is being induced, G's induced group is $-G$, and p^G is the only position equivalent to x.** The negotiating strategy of G, in this case, must be to try to persuade non-members to accept Qx—the future x leads to—as inevitable.

Some further points to note about mutual inducement: **In a multiple, consistent confrontation with two characters, each character must be inducing the other; in one with two groups, each group will be inducing some member of the other; but in other cases, other things are possible. A and B may be inducing each other, while C does as well or better than it would do at its position! Alternatively, A may be inducing B, who is inducing C, who is inducing A.**

Clearly, **an individual in a multiple, consistent confrontation who is not being induced finds x preferred to every position, with the possible exception of its own.**

We assume, in any confrontation, that the interaction between groups has brought them to "cointend" the fixed point x, *in the sense that each* i *intends* x_i *and is aware of the others' intentions and aware that each other is aware*—ad infinitum. It is only because x is cointended in this sense that groups can use it to induce non-members! But if they are so using it, they can't be said to *expect* $x*!*

On the contrary, each group expects—or hopes—that the others will *change* their present positions and intentions and intend its position.

4.6 The inducement paradox

The classic example of inducement occurs in the game of chicken (figure 1). Let row chooser take (4, 2) as its position, and column chooser (2, 4). Their only inducing policies are the "constant" ones of choosing the second row and column, regardless of each other's intention; these are determinate, with a confrontation point at (1, 1).

The problem for either player in chicken is to make the other believe that it truly intends the cell (1, 1) when this intention is irrational! How can a player truly intend a cell that gives it its lowest payoff? Yet it must convince the other of this intention if it hopes to induce the other to accept its position.

"Sure-thing" policies. The general paradox of inducement is that an inducing policy, designed to influence others into accepting a character's position, will often require that character to make credible intentions which are irrational. *How* often it will require this is indicated by a theorem about "sure-thing" policies.

Define a policy for an individual as "sure-thing" if it consists of reacting rationally to every x, i.e., $f_{\{i\}}$ is *sure-thing* if

$$\forall\, x: f_{\{i\}}x \geq_i [x_{-\{i\}}]. \qquad \text{(D44)}$$

By successfully carrying out a sure-thing policy (**which must exist if the frame is finite and ordinal**) i guarantees that, whatever happens, it will have made the best possible choice for itself; it is "sure" to be able to justify its choice, after the event. Hence game theorists generally assume that if a player *can* pursue a sure-thing policy, it *must* do so. The point is, however, that a sure-thing policy typically invites others to exploit you for their benefit! **In any multiple, consistent, two-person confrontation where**, as in chicken, **each position is a strict, strong equilibrium, a sure-thing policy offers the other side *their* position (or an outcome equivalent to it) and induces an outcome worse for you, but better for them, than *your* position.**

In chicken, if each side pursues a sure-thing policy, row chooser induces (2, 4), and column chooser (4, 2).

The inducement paradox in general. In its most general form, the inducement paradox consists in the fact that *a character in a group may have potential improvements from the fixed point that are not equivalent to its position*. In other words, it consists in the existence of a "temptation" for a character $i \in G$ to defect from x to a point not equivalent to p^G. Such a temptation is an outcome y such that

$$x \leq_i y \in Mx - Q^{-1}\, Qp^G. \qquad \text{(D45)}$$

The improvement y constituting such a temptation may be one i can take alone, one it must take in cooperation with others, or one in which it does not have to take any action, but which it may wish to encourage.

In a multiple, consistent confrontation, any such improvement presents it with a dilemma and a corresponding credibility problem. If it is a strict improvement, the dilemma is: *if no one will budge from its intention to implement* x, *should I (irrationally) decide to implement* x *also, or should I try to take, or encourage others to take, an improvement?* The corresponding credibility problem is: *how can I convince others that I will be irrational and reject any improvement?*

If improvements are non-strict, the dilemma and credibility problem are less harsh, but still exist. I have no reason *not* to take such an improvement—indeed, the slightest encouragement would make it irrational to reject it. Therefore, I still have to convince those in our induced group that I will not do so.

If I or others involved in the improvement can't convince them, my group's inducing policy—our attempt to get them to adopt our position—will have no force. Of course, it may be that my improvement too is "inducing," i.e., it may be no better for some non-member of our group than our position. But if I fall back on this argument I in effect abandon our group policy f_G, the confrontation shifts ground, and has perhaps to find a new fixed point; in any case, a new confrontation must be built up within which to establish credibility!

A character is fortunate, in a sense, to be at a fixed point from which it has no improvement: there is no particular reason why this should happen. Hence the inducement paradox is typical of confrontations in general.

The phase 3 inducement paradox. The improvement from x described above would, if taken, be taken at the *dénouement* phase, phase 5. As in the case of the cooperation paradox, there is another version of the inducement paradox which relates not to phase 5, but to the climax phase, phase 3. As with the phase 3 cooperation paradox, it relates to preference, not for an improvement as such, but for another's position.

This phase 5 inducement paradox is discussed by Harsanyi (1977), who uses the term "blackmailer's fallacy" for the argument: *if a blackmailer can cause $1000 worth of damage to its victim, it must be able to extract a ransom of anything up to that amount.* What exactly is wrong with this argument? Is it indeed a fallacy? Harsanyi does not say directly what is wrong with it, but offers a kind of indirect proof that it is fallacious. He points out that if the argument were valid, the victim could equally claim that it must be able to save any amount up to $1000 *off* the ransom, since any positive amount is better for the blackmailer than nothing.

If, however, we assume rationality on the part of the victim, the argument is valid; the blackmailer can be sure of $999 merely by being obstinate. Equally, the victim can be sure of not having to pay more than $1, assuming a rational blackmailer! Something must be wrong, and we submit that it is the assumption of rationality. Neither side, under these circumstances, can assume that the other will be rational. The pursuit of rationality again becomes paradoxical.

Stated generally, the phase 3 version of the inducement paradox is that if i is being induced then this fact undermines its attempt to induce. Whatever it says can in principle be turned against it. The paradox thus confronts any character who is being induced, i.e., who potentially prefers a position not its own to the confrontation point x. Thus i faces a phase 3 inducement paradox if it faces a "temptation" in the form of an outcome y such that

$$x \leq_i y \in \{p^H \mid H \in A - \{G\}\}. \tag{D46}$$

Phase 3 temptations in the two-person case. In chicken, row chooser's temptation to give in and move from the fixed point (1, 1) to column chooser's position (2, 4) is both a phase 3 temptation—row chooser is tempted at the climax to change positions—and a phase 5 temptation—if neither gives in, and (1, 1) has to be implemented, row chooser will, when it comes to the *dénouement,* be tempted to implement row 1 rather than row 2. This is normal. We have the theorem **in a multiple, consistent, two-person confrontation both positions are potential improvements for both characters from the fixed point. Hence each character faces a phase 3 temptation which is also a phase 5 temptation!**

How does this appear in chicken? Row chooser must argue that column chooser should move to (4, 2), because it (row chooser) will never move to (2, 4). But if it will never move, how can it expect column chooser to? Conversely, what reason can it give for column chooser to move that is not a reason for it to? Thus the argument goes at phase 3; meanwhile, each knows that each may defect from x at phase 5, whatever it says now!

Of course, the abstractness and symmetry which make the argument "if me, why not you?" compelling in the chicken game are usually absent in real life; asymmetries and different backgrounds mean that arguments available to one character may not be available to another. The fact remains that in any frame a character's efforts to induce, being based on the assertion that it will not move from the fixed point, are undermined if it is being induced, since this gives it a rational reason to move.

Phase 3 temptations in a three-or-more person confrontation. In a multiple, consistent confrontation with three or more characters, however, a position a character is being induced to accept is not necessarily a potential improvement for it. I and II, for example, may be inducing each other to accept positions which are not potential improvements for them, because moving to them requires the cooperation of III, who, not being induced, prefers x to both of them. **(III necessarily prefers x to both of them if III is not being induced.)**

In the three-or-more person case, therefore, a character may face a phase 3 inducement paradox without facing a phase 5 one.

4.7 Solving the inducement paradox

We saw that positive emotion and changes in the frame that favor other parties play an essential role in solving the cooperation paradox. When it comes to solving the inducement paradox, negative emotion and changes that work against other parties are the essential factors.

The roles of preference change, irrationality, emotion, deceit, disbelief, rational argument in the common interest, the super-character, and morality. In a multiple confrontation, an individual $i \in G$ can solve an inducement paradox by reversing each "temptation" (y, x), where y is either another position or an improvement nonequivalent to p^G. From D45 and D46, i does this by applying to P_i the transformations

$$\mathrm{Rev}(y, x), \text{ for all } y \in \{p^h \mid H \neq G\} \cup (Mx - Q^{-1} Qp^G).$$

It can then rationalize these transformations in order to get rid of intransivities and to ground its new preferences in a proper system of values. **In chicken, applying the transformations to both players' preferences transforms the game into prisoner's dilemma.**

Pressure to make these preference changes comes from being faced with the inducement paradox. If there is little or no preference friction, the changes take place without difficulty. Strong preference friction, however, causes the pressure to erupt in emotion and irrationality, the latter consisting in following new preferences P_i' which have not succeeded in replacing P_i on a permanent, "rational" basis.

The accompanying emotion is positive toward members of G, with feelings of love, loyalty, solidarity, sympathy, good will, etc. It is also positive toward members of G's induced group *when* i *thinks they may abandon their present position and join the group* G. However, in trying to reverse a particular temptation (y, x), i will feel negative emotion—resentment, anger, envy, etc.—toward members of the induced group who share i's preference for y over x *when* i *thinks they will refuse to join* G. The negative emotion has the function of motivating i to want to "punish" them.

Thus, i may have mixed positive and negative emotions toward induced individuals who share with it a preference (y, x) that it has to reverse; such mixed feelings occur when these individuals seem to be hestitating as to whether or not they intend to join G.

However, negative emotions are essential in solving the inducement paradox. They must be present to overcome preference friction.

As in the case of the cooperation paradox, the possibility of deceit and therefore disbelief brings up the *paradox of belief;* as before, this is solved by appeal to evidence and reason rather than just to raw emotion and irrationality; the latter is effective only when implementation of x is immediate, as when a person gets emotional while waving a loaded gun!

In the appeal to evidence and reason, rational arguments in the common interest are constructed based upon the substantive, real-world common interest that both sides have in G's position. In G's case, these are the interests that led G to choose

p^G as its position; in the case of the induced group, they are the interests that make p^G potentially preferred to x. When accompanied by appropriate emotion, these rational arguments become moral: anger and envy are transformed into righteous indignation; ill will and resentment become attributions of blame and attempts to induce guilt. Similarly, feelings of good will, love, etc., are transformed into expressions of mutual praise, high value, and respect. These moral arguments attempt to construct a super-character composed of G together with recruits from the induced group; by making the latter feel guilty and ashamed, the arguments try to make them join with G in pursuing the position p^G. Otherwise—it is made clear by the force of moral indignation—x will be implemented, however much it hurts!

Examples are found in the anger of a parent threatening a child he or she prefers not to punish; in the wrath of God in the Old Testament similarly threatening his people; in the militant, employer-blaming ideology of workers who have to threaten strikes that cause suffering to themselves; in the irrational rage of a hijacker threatening to blow up a plane he or she is on if his/her demands are not met; and so on. These are examples of the general proposition: *negative emotions have the function of making credible threats one would rather not carry out.* Positive emotions toward fellow threateners and recruits assist in this.

Note that negative emotion is a sign of preferring *not* to implement a threat. Thus, paradoxically, it indicates "love"—in the sense of loving preferences, not loving emotion—on the part of the threatener. This is the case with a parent or with God. Hence the advice: don't threaten a child without showing anger, or you give the impression that you enjoy inflicting punishment!

Implementation in stages. In the case of a temptation (y, x) where y is an improvement, the threat to implement x may have to be carried out at a stage of a multistage *dénouement* when the improvement is still a possibility, but the position p^G (for the sake of which the improvement was rejected) is not. Preference change then has to overcome great friction, inasmuch as it needs to be fully integrated and rationalized in order to outlast the circumstances in which it was born. Strong adherence to morality or a creed of revenge—as in the case of the fictional Count of Monte Cristo—can help. On the other hand, sometimes a threat can be irreversibly implemented in a way contingent upon noncompliance with the threatener's position, i.e., there is the chance of a *fait accompli.* To defend itself against this, the recipient needs to make highly credible preference changes now; this is why the prospect of a *fait accompli* makes recipients extremely angry.

Changes in positions, options or cast. If preference friction is too great, emotional pressure may drive characters, instead of changing their preferences, to go back to phase 2 and change their positions and/or the options open to themselves or others. The latter amounts to selecting a new Q from the environment E. The new Q may even contain new characters, as when characters are driven to appeal to higher authority or introduce a "mediator." The creative energy needed to invent or create new threats that, unlike the threat to implement x, can be made credible, comes from the same emotions—anger, resentment, indignation, etc., toward those being induced and feelings of solidarity toward co-inducers or potential recruits—as those which fuel the attempt at preference change.

If an individual changes its position, this may be a matter of giving in and accepting another group's position. Alternatively, it may be a matter of thinking up a creative compromise—i.e., a position that is worse for the individual but better for those it is trying to induce than its current one. Creativity may also lead to thinking up a future that is better for all! It must be a matter of moving to a position that the individual considers more defensible; but the reasons why it considers it to be so may be substantive ones not captured in the abstract model. For example, the new position may fit in better with accepted moral codes.

5. The final state

The order in which we have described paradoxes being overcome need not be taken literally. It is conceptually convenient; it makes an ordered discussion possible; but real-life characters do not have to solve the cooperation paradox first, then the indeterminacy and deterrence paradoxes, and only after that, the inducement paradox.

A group arguing with another at a confrontation may use arguments such as "I can't accept your position because we couldn't all be trusted not to defect from it!" This refers back to the way in which the cooperation paradox has been solved or not solved. It says: "Your position is inconsistent!" Similarly, a group might argue: "Even if I accepted your position, we'd have no way of inducing that other person to accept it." In other words: "Your position is unrealistic—you've no solution to the deterrence paradox." It might even argue: "You say that *x* is no better for me than your position. I'm not sure—*x* could turn out well for me if I can outguess you." In other words, "I don't accept your reification of your policy; you haven't solved the indeterminacy paradox." This argument denies that a properly constructed confrontation—i.e., one with a fixed point—exists.

In other words, recycling takes place between phases 2 and 3 and within phase 3 itself not only because characters reframe the confrontation in new ways, but also because they retract and refuse to accept previous reframings.

Second theorem of the final state. With this warning against a literal interpretation of the order of solution, we have now shown how all the paradoxes of rationality are solved. Where are we?

We have reached the end of the drama—the final state described by Milton (*Samson Agonistes*) as follows:

His servants he, with new acquist
Of true experience from this great event,
With peace and consolation hath dismissed,
And calm of mind, all passion spent.

Our second "theorem of the final state" is: **a confrontation in which no character faces an inducement paradox is united, that is, all characters belong to a single group with a common position.**

On the other hand, we already know (first theorem of the final state) that **if a cast is united and no character faces a cooperation paradox, their common position is a strict, strong equilibrium.**

No nontrivial problem remains. It is true that even in a confrontation with no inducement or cooperation paradoxes, the characters' fixed point may not be equivalent to their single position. Since, however, their position is a strict, strong equilibrium, they will have no hesitation in moving to it! If anyone were to hesitate, it would mean that it had *de facto* changed position—and all would start off again at phase 2!

If no one hesitates to move to its one position, the characters return to phase 2 and go from there to phase 4.

Our two theorems of the final state show that solutions to the two main paradoxes serve different ends. By solving the paradox of cooperation, exemplified by prisoner's dilemma, characters enable themselves to work together as a group in pursuit of agreed objectives. This does not help them, however, if they do not have agreed objectives—if their positions are multiple! This problem they solve by solving the paradox of inducement, exemplified by chicken.

What happens if in phase 3 characters cannot unite behind a single position? The converses of the two theorems are: **if a confrontation is not united, some characters face an inducement paradox. If positions are inconsistent, some face a cooperation paradox.** Since we have hypothesized that paradoxes generate emotional energy and attempts to reframe the situation, the implication is that drama and emotional turmoil will continue.

References

Festinger, L. (1957). *A Theory of Cognitive Dissonance.* Stanford University Press.

Frank, R.H. (1988). *Passions with Reason: The Strategic Role of the Emotions.* New York: W.W. Norton.

Harsanyi, J.C. (1977). *Rational Behavior and Bargaining Equilibrium in Games and Social Situations.* Cambridge University Press.

Howard, N. (1994). "Drama Theory and its Relation to Game Theory. Part 1: Dramatic Resolution vs. Rational Solution," *Group Decision and Negotiation* 3(2).

Howard N. et al. (1993). "Manifesto for a Theory of Drama and Irrational Choice," *Systems Practice* 6(4).