3 Dynamic Deliberation: Stability

> Since in applications of dynamical systems, one cannot pinpoint a state exactly, but only approximately, an equilibrium must be stable to be physically meaningful.
>
> —M. W. Hirsch and S. Smale (1974)

The existence of joint deliberational equilibria corresponds to a consistent solution to the joint decision problem of many players. In the last chapter I discussed conditions under which joint deliberational equilibria exist, under which conditions they correspond to Nash equilibria of a game, and under which conditions they are reachable by Bayesian deliberation starting at a completely mixed point of indecision. The last consideration gives one natural principled motivation for refinements of the Nash equilibrium concept.

Given the strong simplifying assumptions of this discussion, it is natural to raise the question of robustness—of sensitivity to small changes in various aspects of the model. This sort of consideration is a different kind of motivation for refinements of the Nash equilibrium, one aspect of which is dramatized in the metaphor of the "trembling hand." Within the framework of dynamic deliberation, questions of stability and robustness can be categorized and investigated with standard tools of the theory of dynamical systems. In this chapter, I will give some indications of the directions that such analyses can take.

Dynamic Stability of Equilibria

An equilibrium point, *e*, is stable under the dynamics if points nearby remain close for all time under the action of the dynamics.[1] It is strongly stable (or asymptotically stable) if there is a neighborhood of *e* such that the trajectories of all points in that neighborhood converge to *e*. The basin of attraction of a strongly stable equilibrium is the union of all trajectories that converge to it. An equilibrium is unstable if it is not stable. A dynamically unstable equilibrium is the natural focus for worries about the "trembling hand," since in this case there is a neighborhood, N, of e such that for every neighborhood, N', inside N, the trajectory of some point originating in N' leads outside N. Thus, for a dynamically unstable equilibrium, confining the "trembles" to an arbitrarily small N' cannot guarantee that the trajectory stays within N.

Let us reconsider the game of Chicken with the Nash dynamics from this perspective. (I will turn to the effects of varying the dynamics later in this chapter.) The phase portrait is given in Figure 3.1: the lower right corner represents probability one that both players swerve; the upper left represents probability one that neither swerves. The lower left and upper right corners correspond to the two Nash equilibria in pure strategies: Row swerves and Column doesn't; Column swerves and Row doesn't. These are both strongly stable equilibria. The first has as its basin of attraction every point to the lower left of the diagonal and the second has as its basin of attraction every point to the upper right.

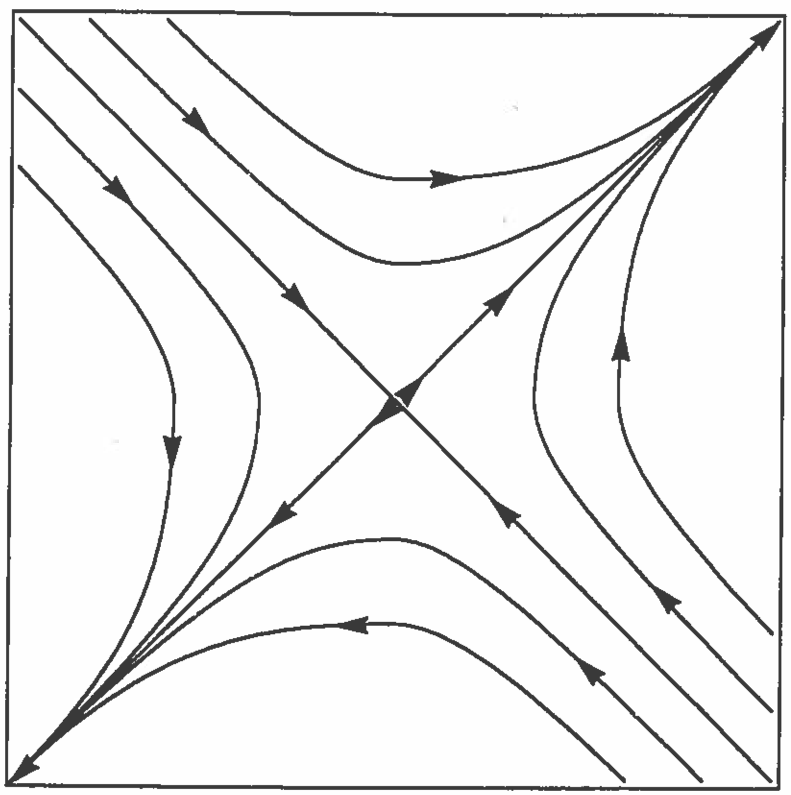


Figure 3.1. The Nash dynamics of Chicken with two stable pure equilibria and an unstable mixed equilibrium

The mixed equilibrium at [0.5,0.5] is unstable. It is called a saddle-point equilibrium, since the dynamics carries every point on the diagonal to it and every point off the diagonal away from it. Notice that this is perfect equilibrium, but it is far from being robust under trembles unless the trembles are somehow conceived as confined to the diagonal.

Pure equilibria can be dynamically unstable. Recall Myerson's game:

Myerson's game

	C1	C2	C3
R3	−9, −9	−7, −7	−7, −7
R2	0,0	0,0	−7, −7
R1	1,1	0,0	−9, −9

Of the three Nash equilibria in pure strategies [R1, C1], [R2, C2], and [R3, C3], both [R2, C2] and [R3, C3] are dynamically unstable under both the Nash and Darwin dynamics while the proper equilibrium at [R1, C1] is strongly stable.[2]

Mixed equilibria can be dynamically stable, and even strongly stable. As an example we can take Matching Pennies under the Nash dynamics.

Matching Pennies

	C1	C2
R2	1,0	0,1
R1	0,1	1,0

Here the unique Nash equilibrium of the game is in mixed strategies at [0.5,0.5]. This is a strongly stable equilibrium under the Nash dynamics,[3] with the whole space as its basin of attraction. Trajectories spiral in as they converge to this point. A typical example is shown in Figure 3.2.

The general phenomenon just described does not depend on there being no Nash equilibria in pure strategies. Consider the following example of Moulin (1986):

Moulin's game

	C1	C2	C3
R3	1,3	2,0	3,1
R2	0,2	2,2	0,2
R1	3,1	2,0	1,3

There is a unique Nash equilibrium in pure strategies [R2, C2], but for each player act 2 is weakly dominated by both act 1 and act 3. If a player does not assign probability one to the other player's doing act 2, his own acts 1 and 3 both look better than his act 2. Consequently, [R2, C2] is highly unstable. Under Nash dynamics, every point in the interior of the space of indecision goes to the mixed equilibrium, where each player gives probability 0.5 of playing act 1 and of playing act 3. The orbits of Row [0.3, 0.7, 0] and Column [0, 0.7, 0.3] are shown in Figure 3.3.

The last two examples point up a difference between a static and a dynamic view of stability. In static discussions of game theory, it is often remarked that mixed equilibria are intrinsically unstable because if your opponent plays the equilibrium strategy you can do just as well by playing any pure strategy with positive weight in your mixed equilibrium strategy as by playing the mixed equilibrium itself. The situation changes if you and your opponent are treated as dynamic deliberators. In this case mixed equilibria may or may not be dynamically stable, and each case must be evaluated on its own merits.

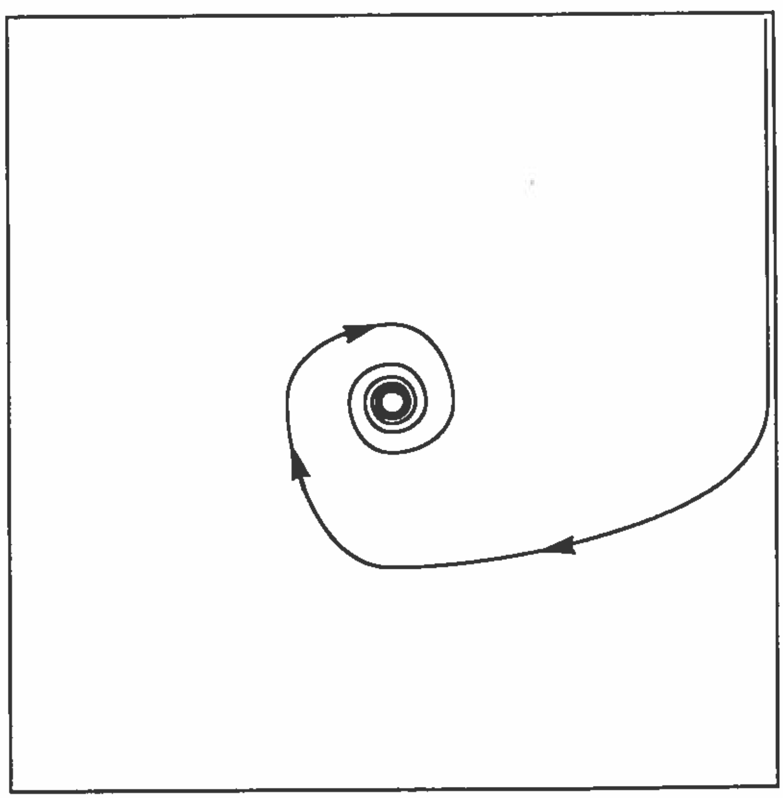


Figure 3.2. The Nash dynamics of Matching Pennies with a stable mixed equilibrium

Imprecise Priors and Elicitation through Deliberation

Our idealized model of games played by Bayesian deliberators makes the unrealistic assumption that at the onset of deliberation precise states of indecision of the players are common knowledge. It is of interest to explore the consequences of weakening this assumption, and it can be weakened in various ways. The prior states of indecision might not be common knowledge or they might not be precise or both. I will postpone discussion of the relaxation of the common-knowledge assumption and concentrate here on imprecise states of indecision.

There are various ways in which imprecise states of indecision might be modeled. Here I will discuss the computationally simplest alternative. Instead of taking a player's state of indecision to be a probability measure over his space of final actions, I will present it as a convex set

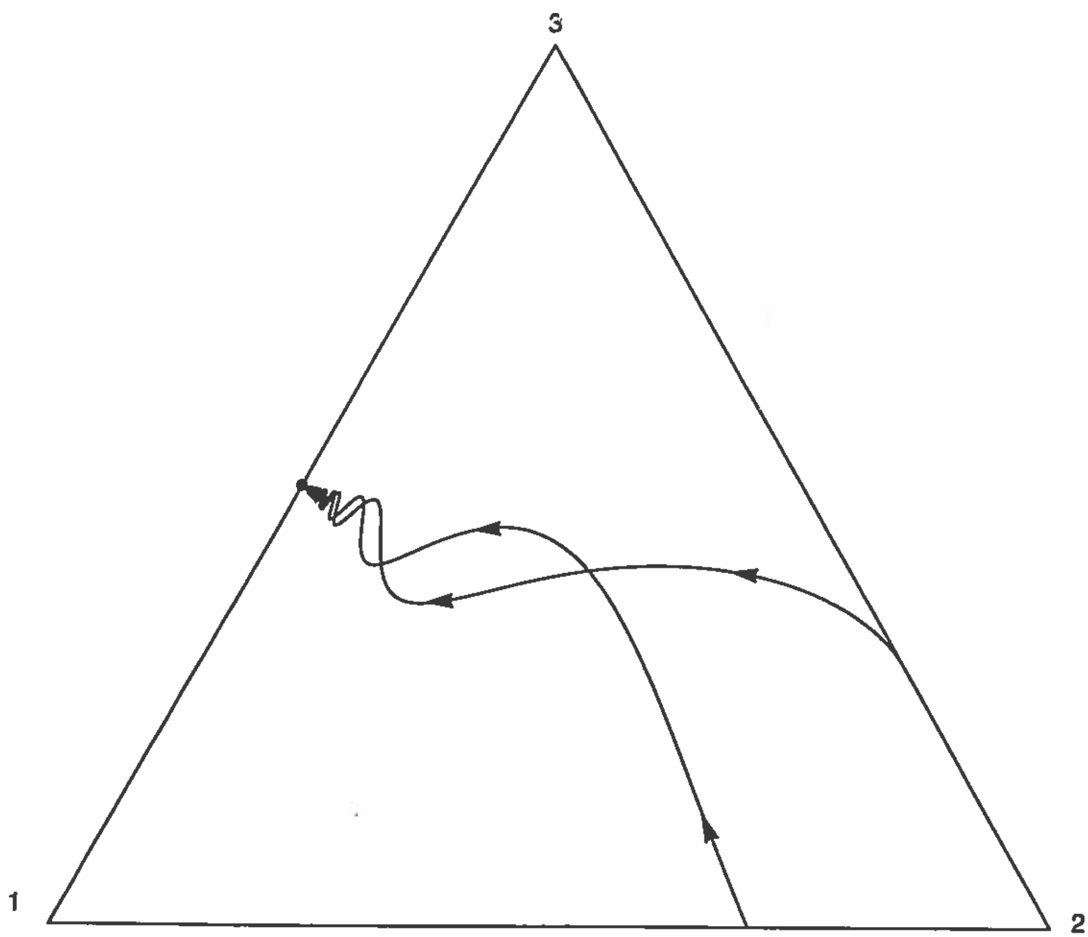


Figure 3.3. Stable mixed equilibrium in Moulin's game

of probability measures (as suggested by Good, 1950; Smith, 1961; Kyburg, 1961; Levi, 1974). I will focus on the simplest case of two-person games, where each player has only two possible actions and a player's state of indecision is given by a closed interval. If, for example, Row's probability of act 2 is to lie in the interval between Row's upper probability of act 2 = 0.7 and Row's lower probability of act 2 = 0.6, then the extreme probability measure corresponding to Row's upper probability of act 2 is $p(\mathrm{A2}) = 0.7$, $p(\mathrm{A1}) = 0.3$, and the extreme measure corresponding to Row's lower probability of act 2 is $p(\mathrm{A2}) = 0.6$, $p(\mathrm{A1}) = 0.4$. The convex set in question is composed of all probability measures over the space [A1,A2] that can be realized as a weighted average of the extreme measures.

How should Row calculate expected utilities given Column's probability interval? He should have a set of expected utilities, one corresponding to each possible point probability consistent with Column's probability interval. Because of the nature of the expectation, however, Row need only compute the expected utilities relative to the endpoints of Column's interval with assurance that the other point utilities lie between the endpoints.

How should Row modify his probability sets in the light of new expected utility sets? He should have new probability sets corresponding to every point reached by applying his dynamical law to a point chosen from the expected utility set and a point chosen from his old probability set. But for the Nash dynamics, and a large class of reasonable dynamical laws to which it belongs, it is a consequence of the form of the dynamical law that if the old probability of an act, A, is in the interval between the upper and lower probabilities of the act [call the two extremes $p_1(A)$ and $p_3(A)$ and the point in between $p_2(A)$], and if the old utility of the act is in the interval of utilities corresponding to the probability interval [call the utilities $U_1(A)$, $U_2(A)$, and $U_3(A)$], then the new probability, p', is in a new probability interval [that is, $p'_2(A)$ is between $p'_1(A)$ and $p'_3(A)$]. It is a consequence of these observations that Row can achieve the results of point deliberation on every pair consisting of one point from his interval and one from Column's interval by performing four point computations on pairs consisting of one endpoint from his interval and one from Column's interval. The new maximum and minimum probabilities of A among the four possibilities form the endpoints of his new probability for A. The general points made above continue to hold good *mutatis mutandis* for numbers of acts greater than two, with intervals being generalized to convex sets of probability measures and endpoints being generalized to extreme points. With regard to computational tractability, deliberational dynamics, as so far developed, has a certain affinity for convex set representations of imprecise probabilities.

In the case of two players each of whom must choose between two acts, a state of indecision in the interval-valued sense is now represented as a rectangle in the old space of indecision—the product of Row's and Column's intervals. Points are considered degenerate intervals, and point states of indecision are special cases of rectangles of indecision. At the other extreme a player's interval may be the whole interval [0,1], in which case we will say that he is totally bewildered, and where the state of indecision is the whole space, we will say that the players are in a state of mutual total bewilderment. A rectangle of indecision that the dynamics maps onto itself is a dynamical equilibrium state.

The area of a rectangle of indecision need not be preserved by deliberational dynamics. For example, players may start out with nondegenerate interval-valued probabilities and be carried by deliberation to point probabilities. One might call such a process elicitation of point probabilities through deliberation. It is illustrated in Figure 3.4 in terms of the

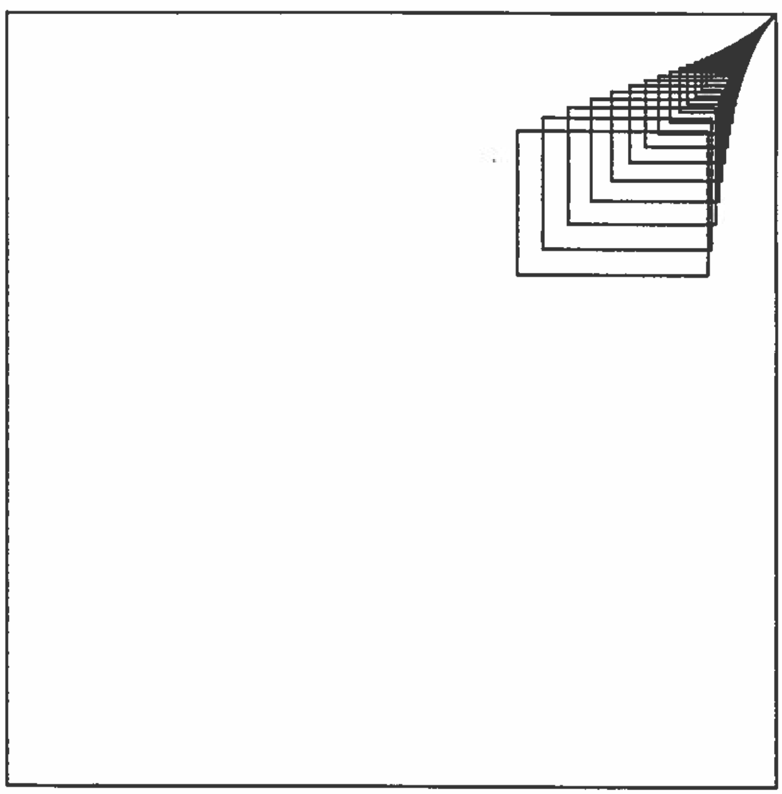


Figure 3.4. The Winding Road with interval-valued probabilities

Nash dynamics applied to a pure coordination game, The Winding Road. Row starts with a probability interval of [0.6, 0.8] and Column starts with a probability interval of [0.6, 0.9]. They each converge to a point probability of one (of driving on the right).

The same process in the case of a game with elements of both competition and coordination is illustrated in Figure 3.5. Here, in the game of Chicken, we have the orbit of [0.4,0.1], [0.4,0.1] converging to [0,0] and that of [0.9,0.6], [0.9,0.6] converging to [1,1]. The effect is illustrated in the extreme in Figure 3.6, where an initial state of mutual total bewilderment is carried to the point equilibrium [Defect, Defect] by the Nash dynamics in Prisoner's Dilemma.

It is evident that much of our analyses of these games is robust under generalization to interval-valued probabilities. Let us look at the matter a little more closely. Let us say that a point equilibrium is here robust under imprecision if there is a nondegenerate rectangle of indecision that contains the point and converges to it. Figure 3.6 shows that [Defect, Defect] in Prisoner's Dilemma is robust under imprecision. [Right, Right] and [Left, Left] in The Winding Road and [Row swerves,

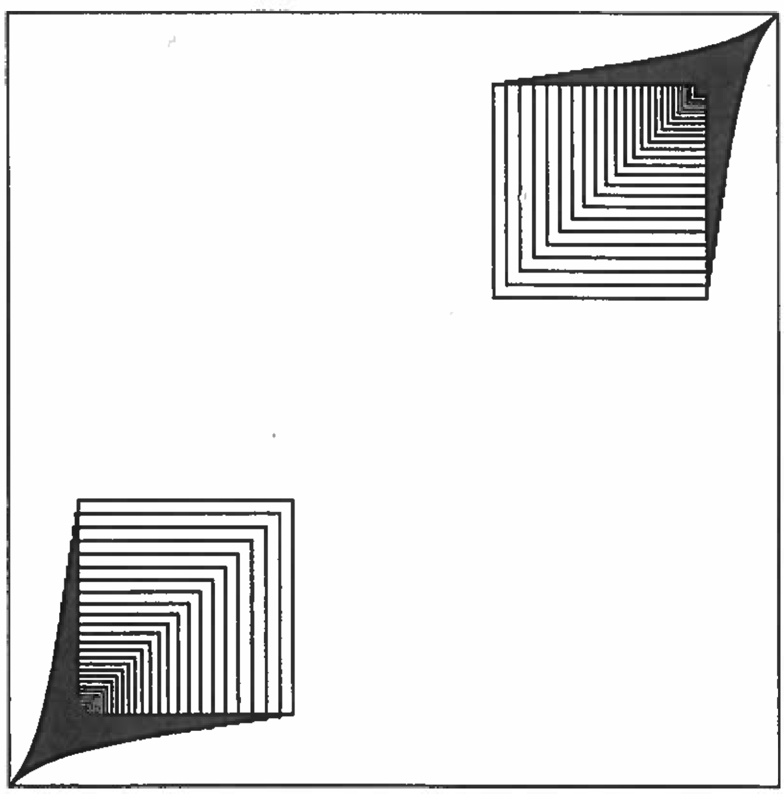


Figure 3.5. Chicken with contracting intervals

Column doesn't] and [Column swerves, Row doesn't] in Chicken are all robust under imprecision. However, the mixed equilibria in these games are not. For example, consider the mixed point equilibrium at [0.5,0.5] in Chicken. The orbit of the small rectangle, [0.51,0.49], [0.51,0.49], explodes to a state of mutual total bewilderment, as does any rectangle that straddles the separatrix of the point dynamics (the diagonal from upper left to lower right). The orbits of nondegenerate rectangles that include the mixed point equilibrium but do not straddle the separatrix suffer a more modest explosion, as shown in Figure 3.7, where an initial rectangle of [0.5,0.4], [0.5,0.4] is carried to an equilibrium rectangle of [0.5,0], [0.5,0].

In the examples given so far the point equilibria that were robust under imprecision were ones which were strongly stable in the point dynamics. One might suspect that these notions coincide, but this conjecture is shown false by the simple example of Matching Pennies. Recall that this game has only one Nash equilibrium point at [0.5,0.5] and that for point probabilities under the Nash dynamics this point is a

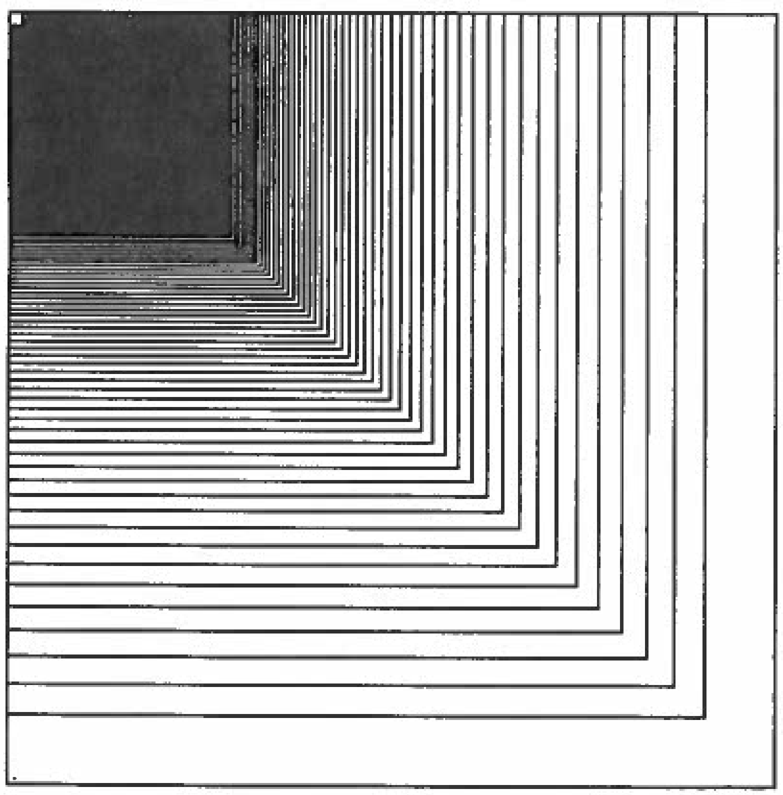


Figure 3.6. Prisoner's Dilemma with contracting intervals

strongly stable equilibrium that is an attractor for every point in the space. If, however, we start with nondegenerate rectangles rather than points in the joint space of indecision, the situation is reversed. Every nondegenerate interval explodes to a state of mutual total bewilderment. This is illustrated in Figure 3.8 for an initial rectangle of [0.51,0.49], [0.51,0.49]. Other nondegenerate rectangles do no better. Robustness under imprecision is a stronger variety of stability for equilibrium points than strong stability is in the point dynamics.

The analysis of Matching Pennies for point states of indecision does not deserve to be called robust under imprecision. What about our analyses of The Winding Road and Chicken? In each of these cases, the pure Nash equilibria are robust under imprecision. Any rectangle that does not touch the separatrix diagonal has an orbit that converges to one of the pure Nash equilibrium points. On the other hand, the mixed Nash equilibrium point is not robust under imprecision and interaction with the diagonal leads to trouble. How much of the space leads to trouble depends on how imprecise the players' priors are. One can get some

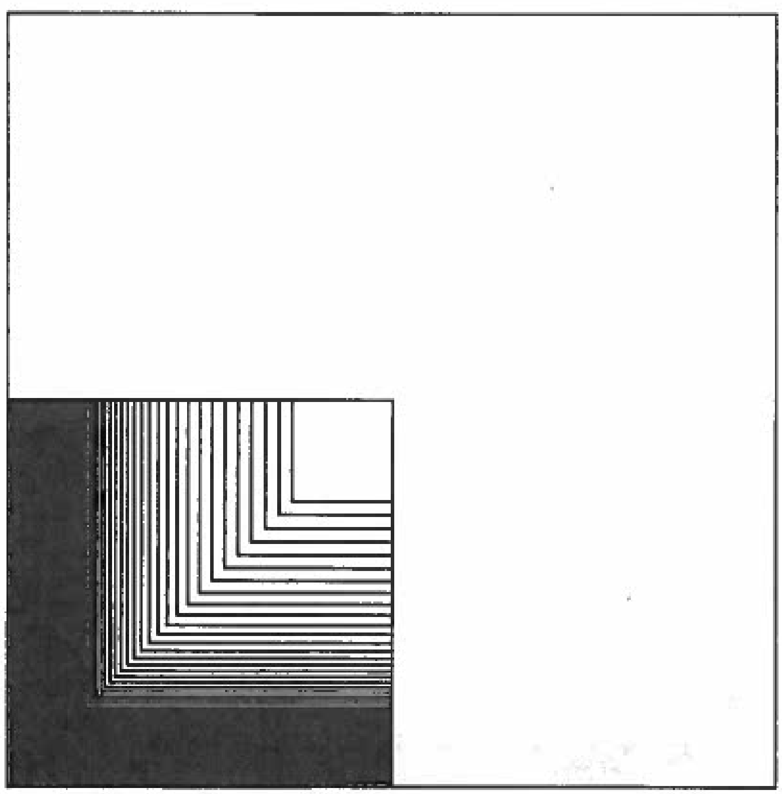


Figure 3.7. Chicken with expanding intervals

idea of the magnitude of the difficulties by putting a grid over the space of indecision. In a regular grid with 0.1 spacing, ten of the squares straddle the diagonal and eighteen more touch it, for a total of 28 percent troublemakers. For a spacing of 0.05 about 15 percent are troublemakers, and for a spacing of 0.01 the proportion of troublemakers drops to about 2 percent.

When analyzed in terms of point priors, The Winding Road and Chicken were both seen to be situations in which coordination could arise spontaneously. In fact, in that setting, it seemed to require a miracle for coordination *not* to occur. The conclusion that coordination can occur spontaneously in such situations continues to hold good for imprecise priors. But the modeling of this section, although still far from realistic, does give us some reason to expect trouble near the diagonal. (Exactly what sort of trouble it is depends on what, if anything, bewildered players are supposed to do. See the discussion of interval-valued probabilities in Chapter 5.) Coordination does not appear quite so effortless, and it would be to everyone's mutual advantage to set up the pre-

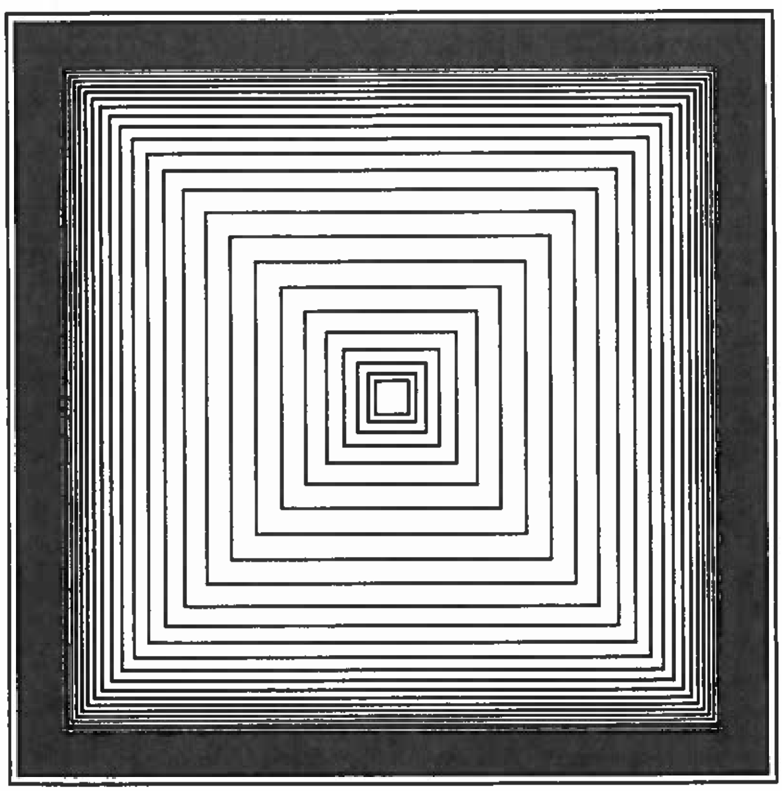


Figure 3.8. Intervals explode in Matching Pennies

deliberational environment to keep players away from the diagonal. For The Winding Road, a liberal use of street signs might do the trick. Or consider The Intersection, a sort of attenuated game of Chicken discussed in Moulin (1986):

The Intersection

	Go	Stop
Go	−10, −10	1, −.1
Stop	−.1,1	0,0

A stoplight in the predeliberational environment is a reasonably effective means for keeping the players away from the diagonal.

Analysis of robustness under imprecision could be developed further, and imprecision could be modeled in other—perhaps more realistic—ways. In what follows I will for the most part deal with simpler models with point probability states of indecision, but I wish to point out that additional questions of robustness under imprecision are always in order.

Structural Stability I

Another kind of dynamic stability is of interest. That is the stability of the location and type of equilibrium points as differential equations are varied. One way of varying the differential equations keeps the fundamental dynamical law the same but varies the payoffs of the game. I will illustrate with game-theoretic models of the arms race. Political philosophers often model the arms race as Prisoner's Dilemma, with the following payoffs:

Prisoner's Dilemma

	Defect	Don't defect
Defect	−5, −5	5, −10
Don't defect	−10,5	0,0

Deliberational dynamics inexorably carries every point in the space to the tragic strong equilibrium of both sides deciding to arm (Figure 3.9). But, at least in some arms races, some generals and some politicians may think that the proper model of the arms race is not Prisoner's Dilemma but, rather, Chicken. The disagreement is about the relative values of D and R in the payoff matrix:

	C1	C2
R2	D,D	$5,R$
R1	$R,5$	$0,0$

With $D = -5$ and $R = -10$ we have Prisoner's Dilemma; with $D = -10$ and $R = -5$ we have Chicken. With $D = R = -10$ we have a structurally unstable transition game, which has the following payoff matrix and which is plotted in Figure 3.10.

	C1	C2
R2	$-10, -10$	$5, -10$
R1	$-10, 5$	$0, 0$

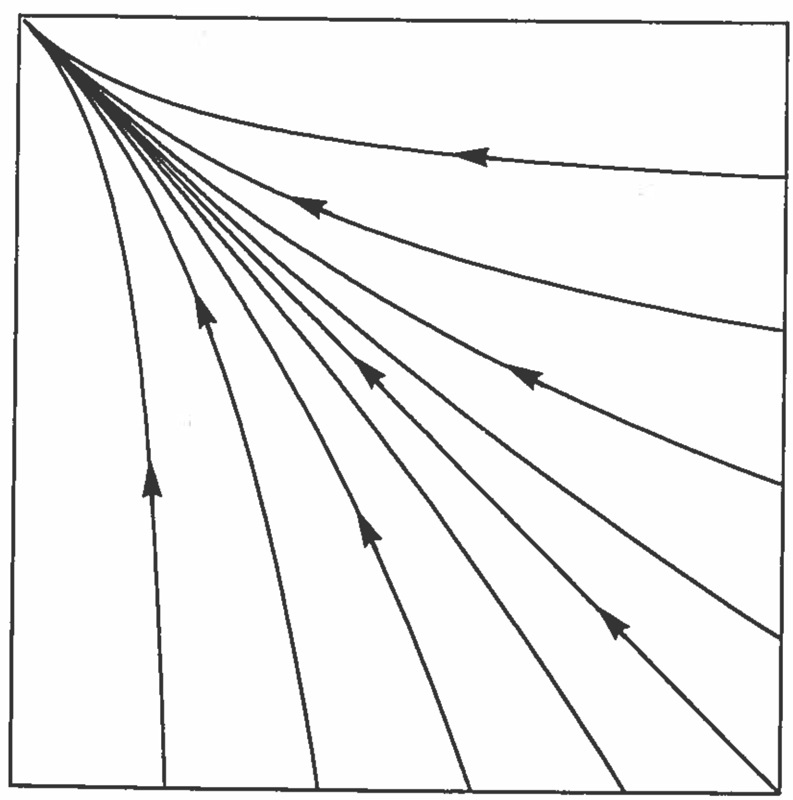


Figure 3.9. Nash dynamics of Prisoner's Dilemma

Figure 3.10 looks much like the portrait of Prisoner's Dilemma in Figure 3.9, but there are some subtle changes: there are additional (unstable) equilibria in pure acts at [0,0] and [1,1] and additional (unstable) equilibria in mixed acts at $x = 0$, $y < 1$ and at $y = 1$, $x > 0$. These equilibria are indicated in the figure by the bold lines. The equilibrium at [0,1] is still stable, but it is no longer strongly stable for it is not an attractor for orbits of the aforementioned mixed equilibrium points. If D is allowed to creep a little below R, then we have The Birth of Chicken (Figure 3.11), whose payoff matrix is:

The Birth of Chicken

	Don't swerve	Swerve
Don't swerve	$-10.5, -10.5$	$5, -10$
Swerve	$-10, 5$	$0,0$

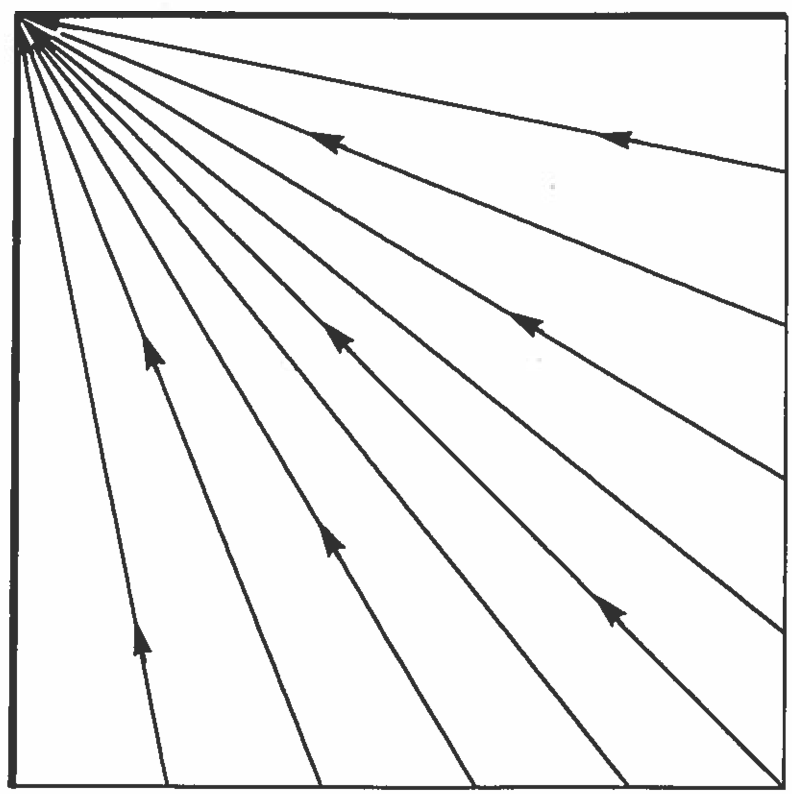


Figure 3.10. Transition from Prisoner's Dilemma to Chicken

There is now a dramatic change. The equilibrium point formerly at [0,1] moves down the diagonal and changes from a stable equilibrium to a hyperbolic one. The equilibria at [0,0] and [1,1] change from unstable to strongly stable. They are now attractors for the orbits of almost all points in the space. The former mixed equilibria on $x = 0$ and on $y = 1$ have vanished. We have passed through the "better *R* than *D*" bifurcation. The equilibrium points as a function of decreasing *D* are plotted in Figure 3.12.

Other transitions are of interest. Consider a Dove's model of the arms race. The Dove may well believe that the payoff in the case of mutual disarmament is greater than the payoff in the case in which her country arms and the other doesn't. After all, arming diverts economic resources and may tempt her own country's political leaders into adventures they had best not undertake. Thus, she thinks that in the payoff matrix

	C1	C2
R2	*D*,*D*	5,*R*
R1	*R*,5	*P*,*P*

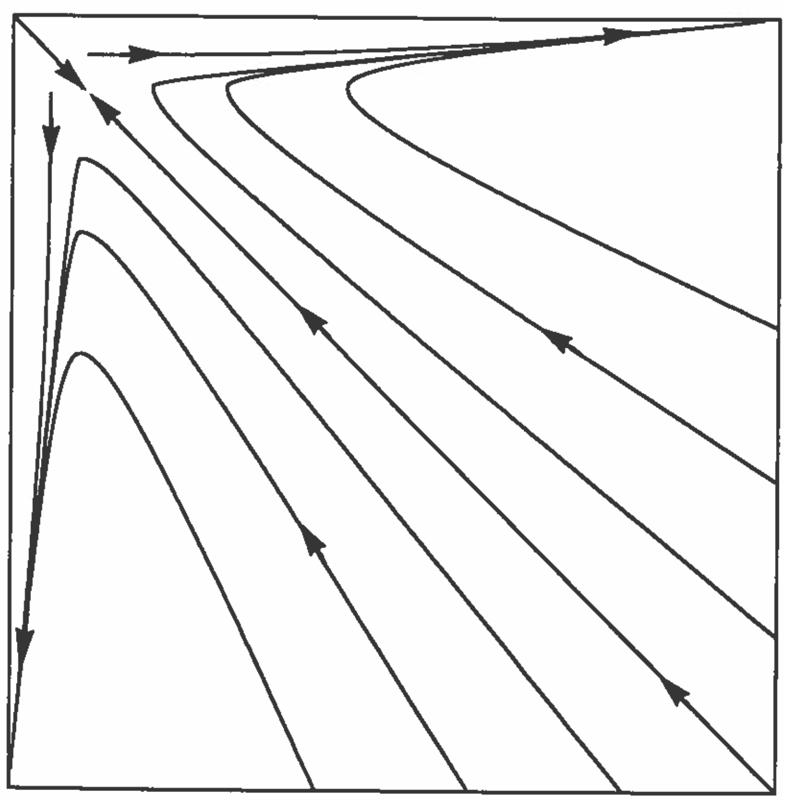


Figure 3.11. The Birth of Chicken

P should be greater than 5. Figure 3.13 shows the transition from Prisoner's Dilemma to the Dove model with $D = -5$, $R = -10$, $P = 5.5$. A tiny window of hope has opened in the lower right corner. It is bounded by a saddle-point equilibrium on the [0,1]–[1,0] diagonal. Orbits of points in its interior are attracted to a new stable equilibrium in the lower right corner. If the players come to this game with enough prior goodwill to put the initial point in the window, they will be carried to this equilibrium. This Dove game is also a test of will.

All of these games (and other variations that may suggest themselves to you) may be more or less reasonable models for arms races in various specific situations. For those who would model the arms race in this way it is of real interest to investigate the dramatic changes in deliberational equilibria that can occur as the payoffs are continuously varied. In the first place, doing so would give important information as to the robustness of the model. And if the model is accurate, neighborhoods of structural instability may represent situations of great risk or great opportunity.

One useful concept of structural stability was introduced into the

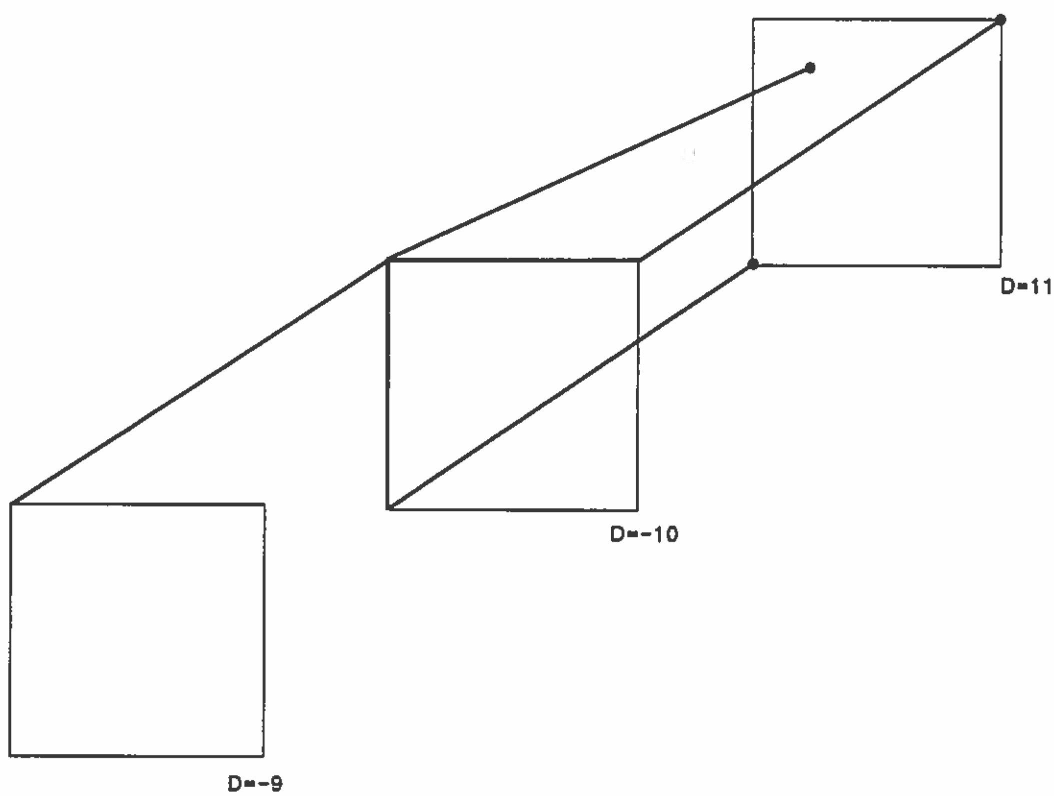


Figure 3.12. The "better R than D" bifurcation

game-theoretic literature by Kohlberg and Mertens (1986), who call a Nash equilibrium of a game hyperstable if it is robust under small variations in the payoffs. Hyperstability can be studied without reference to the particular dynamics since it depends only on how the Nash equilibria move, and Kohlberg and Mertens did not consider dynamics.[4] But from the point of view of deliberational dynamics, hyperstability can be viewed as a concept for classifying joint deliberational equilibria.

More precisely, a Nash equilibrium, N, is hyperstable if for every ε there is a δ such that if the pure strategy payoffs are perturbed by less than δ there is a Nash equilibrium in the open ball with radius ε centered at N. Not every game has a hyperstable equilibrium, but almost all do. That is, except for a set of Lebesgue measure zero in the space of pure strategy payoffs, every normal-form game has a hyperstable equilibrium. If a game has only a finite number of equilibria, then one of them is hyperstable. This sort of structural stability is not unrelated to previous concerns. Hyperstable equilibria are always proper and perfect (see Kohlberg and Mertens, 1986; Leal, 1986).

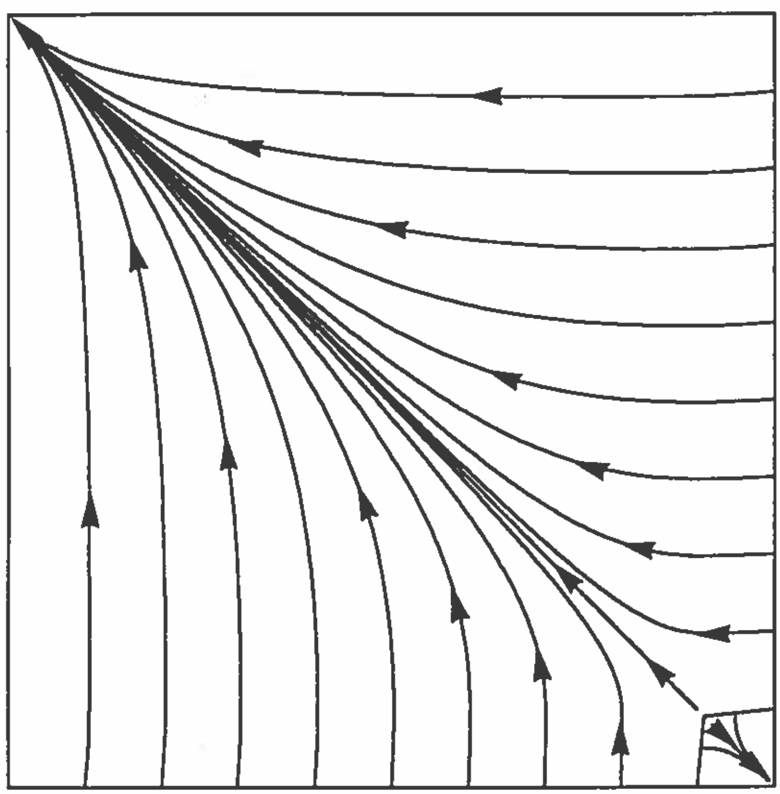


Figure 3.13. The Birth of Dove

For example, consider Samuelson's game, in which Darwin can converge to imperfect equilibria:

	C1	C2
R2	11,11	11,11
R1	12,10	9,8

The imperfect equilibria at $p(\text{R2}) = 1$, $p(\text{C2}) > 0.25$ are not hyperstable, because they cease to be equilibria in the perturbed game:

	C1	C2
R2	$11{,}11 + \varepsilon$	11,11
R1	12,10	9,8

The perfect equilibrium at $p(\text{R1}) = p(\text{C1}) = 1$ is hyperstable. In fact, small perturbations in the pure strategy payoffs do not move it at all.

For a somewhat different kind of example, let us reconsider Matching Pennies. Here there is only one Nash equilibrium: the mixed strategy where each player gives each of his pure strategies probability 0.5. This equilibrium is hyperstable. Here small perturbations in the payoffs can move the equilibrium, but arbitrarily small perturbations move the equilibrium an arbitrarily small amount.

Considerations of hyperstability, however, tell only part of the story about the structural stability of equilibria. Recall the transition game between Prisoner's Dilemma and Chicken:

	C1	C2
R2	−10, −10	5, −10
R1	−10,5	0,0

The Nash equilibria consist of all points with $p(\text{C1}) = 1$ and all points with $p(\text{R2}) = 1$. Consideration of perturbations in the direction of Prisoner's Dilemma and of Chicken is sufficient to show that equilibria other than $E = [p(\text{R2}) = 1, p(\text{C1}) = 1]$ are not hyperstable. In the direction of Prisoner's Dilemma, E does not move; in the direction of Chicken it moves a little bit in response to small perturbations. However, this leaves out the fact that E changes its dynamic stability status, under the Nash dynamics, from strongly stable in Prisoner's Dilemma to stable but not strongly stable, in the transition game, to an unstable saddle point in Chicken. Furthermore, consideration of asymmetric perturbations shows that E is not hyperstable either. For any positive ε, the perturbed game given by

	C1	C2
R2	$-10, -10 + \varepsilon$	$5, -10$
R1	$-10 + \varepsilon, 5$	$0,0$

has a unique Nash equilibrium at $p(\text{R1}) = p(\text{C1}) = 1$. So the short story regarding structural stability for the transition game is that there are no hyperstable equilibria but, as we have seen, there is a much richer long story to be told.

Structural Stability II

One can investigate structural stability at a more radical level. Instead of simply changing the payoffs, one can change the dynamical law. Then one can investigate which features of the dynamics are robust under changes of the dynamical laws. For example, let us look once more at Matching Pennies. In the Nash flow:

$$\frac{dp(A)}{dt} = \frac{\text{cov}(A) - p(A)\Sigma_i \text{cov}(A_i)}{1 + \Sigma_i \text{cov}(A_i)}$$

In the closely related Brown–von Neumann flow:

$$\frac{dp(A)}{dt} = \text{cov}(A)^2 - p(A)\Sigma_i \text{cov}(A_i)^2$$

And in the Darwin flow:

$$\frac{dp(A)}{dt} = p(A)\frac{EU(A) - EU(SQ)}{EU(SQ)}$$

The unique mixed equilibrium at $[p(\text{R2}) = 0.5, p(\text{C2}) = 0.5]$ is a strongly stable spiral attractor, having the whole space as its basin of attraction. If we move to the Aristotelian flow for 2×2 games (Skyrms, 1986), such that

$$\frac{dp(\mathrm{A2})}{dt} = EU(\mathrm{A2}) - EU(\mathrm{A1})$$

the vector field changes character, as illustrated in Figure 3.14. The Nash equilibrium is still stable, but it is no longer strongly stable. The deliberators jointly form a harmonic oscillator. The closed orbits are not structurally stable.[5] Slight variations in the dynamics can turn them into outward or inward spirals.[6]

Some games are more sensitive to changes in the dynamics than others. Consider a game like Chicken. The vector field looks qualitatively the same as that illustrated in Figure 3.1 for a wide variety of dynamical rules.[7] The curvature of the orbits may vary, but the destinations of the points remain the same. Every point to the lower left of the diagonal goes to the Nash equilibrium at the lower left corner; every point to the upper right goes to the upper right corner. Every point on the (anti) diagonal (except perhaps endpoints) goes to the saddle-point equilibri-

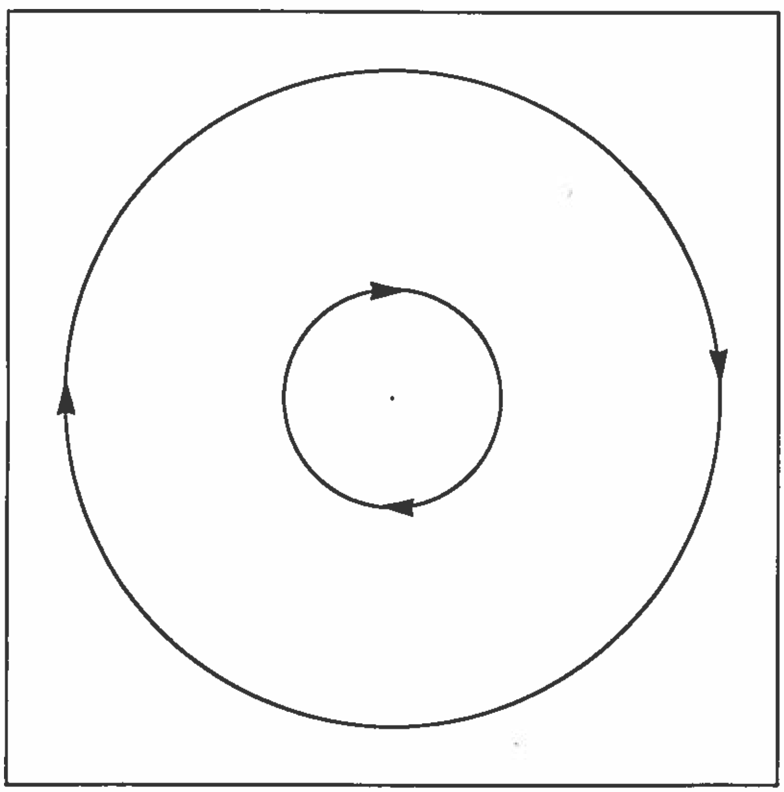


Figure 3.14. Matching Pennies with Aristotelian dynamics

um at [0.5,0.5]. We can argue that any reasonable autonomous dynamics will give this sort of picture.

If the dynamics seeks the good then the qualitative direction of the flow in the four quadrants of the space of indecision is as indicated in Figure 3.15. Suppose $dp(A_i)/dt$ is a continuous function of the expected utilities and probabilities of the A_i, which assumes the value zero only if $U(A_i) = U(SQ)$ or $p(A_i) = 1$ or 0. Consider a point, p, in the interior of the lower left quadrant, as in Figure 3.16. The point must move into the rectangle *APBE* and cannot ever get out. Draw a horizontal line *DC* at $p(\text{R1}) = 1 - \varepsilon$. Within the rectangle *APCD*, $dp(\text{R1})/dt$ is always positive since on *APBE* it is zero only at *E*. By continuity (a continuous function defined on a compact set assumes a maximum and a minimum), $dp(\text{R1})/dt$ is bounded away from zero on *APCD*. So in some finite amount of time the trajectory of p moves into *DCBE*, from which it cannot escape. A similar argument with respect to $dp(\text{C1})/dt$ gets the point to within ε of $p(\text{C1}) = 1$. Since ε is arbitrary the trajectory must converge to the equilibrium, *E*. The situation in the upper right quadrant is similar. The equilibria [Row swerves, Column doesn't] and [Column swerves, Row doesn't] are thus as stable as you please in all the ways we have discussed.

Suppose that a point is on the diagonal from upper left to lower right. Then by symmetry, $dp(\text{R2})/dt = dp(\text{C1})/dt$, and by a continuity argument similar to the one above every point except endpoints must converge to

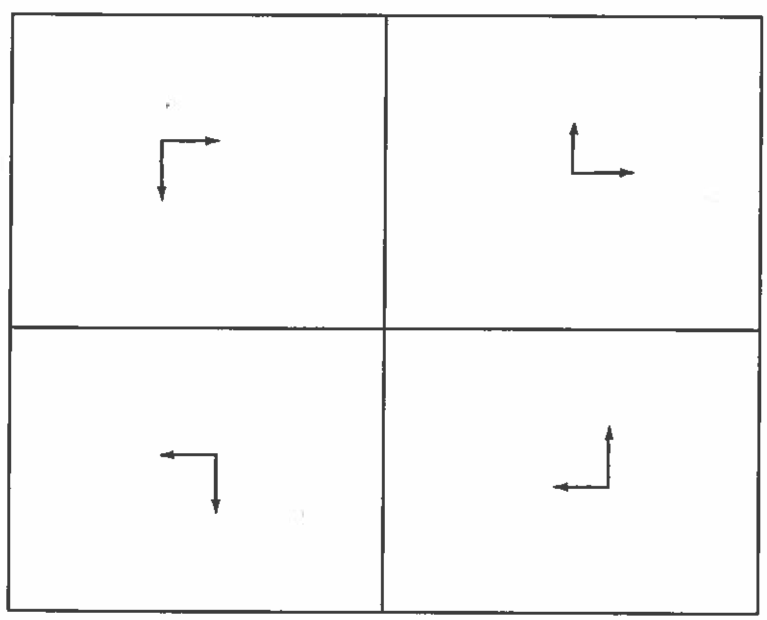


Figure 3.15. Four quadrants in Chicken

the mixed Nash equilibrium at [0.5,0.5]. The situation in the rest of the upper left and lower right quadrants is only slightly more sensitive. Consider the triangle consisting of the space below the diagonal in the upper left quadrant. Drop a vertical line from the diagonal, as shown in Figure 3.17. On the diagonal, $dp(\mathrm{R1})/dt = dp(\mathrm{C2})/dt > 0$. At $p(\mathrm{R1}) = 0.5$,

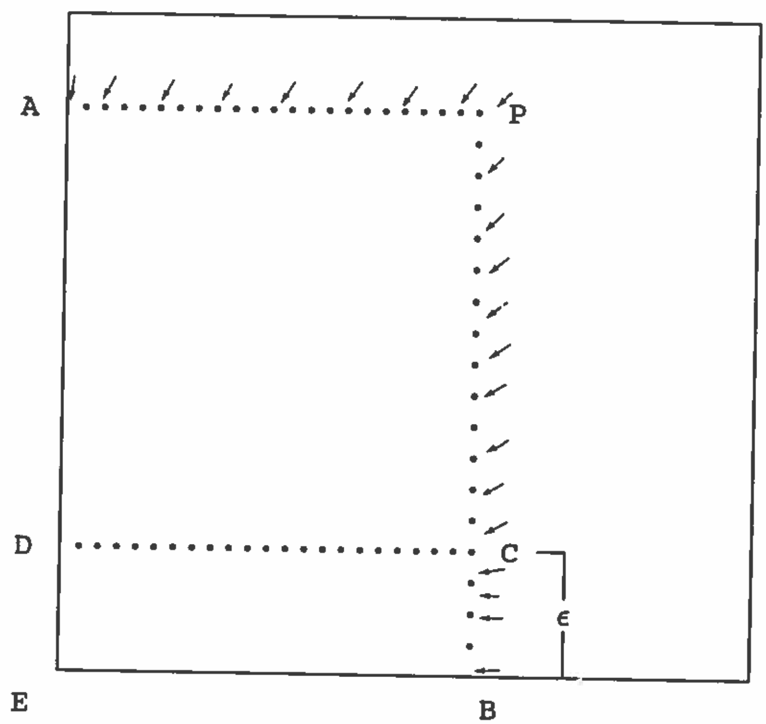


Figure 3.16. Qualitative analysis of Chicken

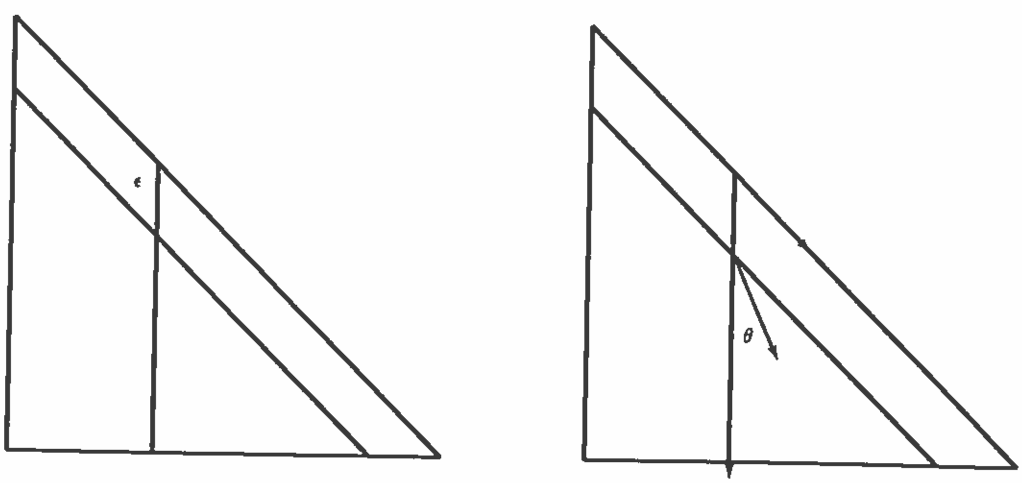


Figure 3.17. Below the diagonal in the upper left quadrant

$dp(\text{C2})/dt = 0$ while $dp(\text{R1})/dt$ is positive. Let us make the additional assumption that the direction of the velocity vector as measured by the indicated angle θ varies continuously and monotonically along this line. Then, for any point in this region, we can draw a line between it and the diagonal such that at that line the vector field always points inward. Since the downward velocity $dp(\text{R1})/dt$ is positive throughout the smaller triangle thus formed, it must by continuity be bounded away from zero and in some finite time the trajectory of the point must emerge into the lower left quadrant, and so be delivered to the lower left equilibrium. The remaining cases are similar. Thus the whole qualitative analysis of Chicken is extremely robust over variations in the deliberational dynamics.

However, there is one sense in which Prisoner's Dilemma is even more robust than Chicken. In the discussion of Chicken, we allowed the dynamical laws to vary among a wide class but we assumed that in any particular situation whatever dynamics was being used was common to the players and common knowledge between them. This extends even to the index of caution. If Row is a Nash deliberator and Column a Darwin deliberator, or even if both are Nash deliberators but one has a greater index of caution, and if these facts are common knowledge, then orbits that originate on the diagonal will not stay on it.

In Prisoner's Dilemma, however, it doesn't matter. Suppose two players have two different dynamical rules from the class of rules which seek the good such that $dp(A_i)/dt$ is a continuous function of the expected utilities and probabilities of the A_i, which assumes the value zero only if $U(A_i) = U(SQ)$ or $p(A_i) = 1$ or 0, and suppose that it is common knowledge that the players have these rules. Then every completely mixed state of indecision converges to [Defect, Defect].[8]

This is a consequence of the fact that [Defect, Defect] is the unique rationalizable strategy. For each player, Defect strongly dominates Cooperate. So, at each point in the space the velocity toward Defect is positive for each player. Then, by an argument like that used regarding the lower left quadrant of Chicken, the players will converge to [Defect, Defect].

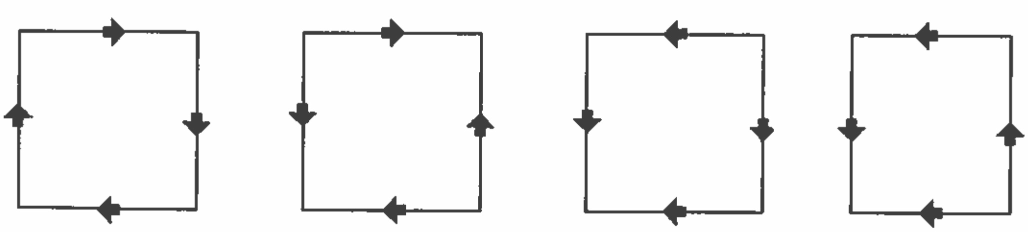
In a reversal of the analogy, the pure equilibria of Chicken each have a neighborhood in which the vector field looks qualitatively like that of Prisoner's Dilemma. They are *strong equilibria* in the sense of Harsanyi (1973a). At the equilibrium each player's strategy is her unique best reply. By continuity of expected utilities as a function of probabilities, a Harsanyi-strong equilibrium should have a neighborhood within which each player's equilibrium strategy is still her unique best reply. Then it should be strongly dynamically stable in any reasonable dynamics. By continuity of expected utilities as a function of utilities of pure strategies, it should be hyperstable. In fact, Harsanyi-strong equilibria appear to possess all the local stability properties that one might desire.[9]

Our three examples of this section illustrate common qualitative patterns in 2 × 2 two-person games. If we consider the space of all such games, as defined by their payoffs, the cases where two payoffs in the game matrix are exactly equal has Lebesgue measure zero. If we neglect these, there are—up to symmetry—essentially the four qualitative situations depicted in Figure 3.18. The first situation is illustrated by Matching Pennies, the second by Chicken and Battle of the Sexes, and the third by Prisoner's Dilemma.[10] We haven't met an example of the fourth situation, but it is much like Prisoner's Dilemma in that the game is solvable by iterated elimination of strictly dominated strategies. C1 strictly dominates C2. When C2 is eliminated, R2 strictly dominates R1. Thus the unique Nash equilibrium is the unique rationalizable strategy.

I do not want to claim that games with Lebesgue measure zero have probability zero of occurring in the real world. There are reasons why we do have games where some payoffs equal others, and so such games are of practical as well as theoretical interest. But the foregoing at least gives some reason to believe that the examples to which we have devoted so much attention are not completely atypical.

Stability and Rationality

In Chapter 2 we saw how the Nash equilibrium concept arises naturally in the context of games played by bounded Bayesian deliberators. A refinement of the Nash equilibrium concept also arises naturally, that of a joint deliberational equilibrium to which deliberation starting in a completely mixed state of indecision can converge. This concept is related to the notions of perfect and proper equilibrium discussed in the game-theoretic literature, but it is not identical with either. It does not depend



1 2 3 4

Figure 3.18. Typical 2 × 2 games

for its interest on any irrationality on the part of the players, although it does depend on their limited computational resources.

In this chapter we surveyed various types of stability and robustness that arise naturally in the context of the dynamics of rational deliberation. These are related to game-theoretic notions of "strategic stability," which are often discussed in terms of a little irrationality on the part of the players. But I think that the types of stability we have surveyed can also be motivated without any presumed irrationality. In Chapters 4 and 5 I will examine just what rationality in these contexts entails.

In a situation where deliberation costs something, deliberation will often be terminated close to but short of a Nash equilibrium. The first theoretical question arising from this consideration concerns the dynamic stability of deliberational equilibria. It is at least arguable that interval-valued priors do not entail any irrationality (see Smith, 1961; Good, 1950; Levi, 1974; Kyburg, 1961), in which case robustness under imprecision is of interest even when rationality is common knowledge. Uncertainty and/or imprecise knowledge about the payoffs of the game is sufficient motivation for concern with the first kind of structural stability we have discussed. Structural stability under variations of the dynamical law are of obvious interest, since in our framework there is no unique rational dynamic law.

There are many interrelations between the various types of stability flowing from the deliberational dynamics, and between these and the many kinds of refinement of the Nash equilibrium that have been introduced in the game-theory literature. Some of these have been pointed out along the way, but we are far from having the whole story.[11] It appears that the framework of dynamic deliberation not only provides a rationale for the concerns of classical game theory, but also suggests fertile areas for new investigations.