
Rationalizability and Common Knowledge of Rationality

Men tracht un Got lacht
 (Mortals scheme and God laughs)
 Yiddish proverb

To determine what a rational player will do in a game, eliminate strategies that violate the canons of rationality. Whatever is left, we call *rationalizable*. We show that rationalizability in normal form games is equivalent to the iterated elimination of strongly dominated strategies, and the epistemological justification of rationalizability depends on the *common knowledge of rationality* (Tan and Werlang 1988).

If there is only one rationalizable strategy profile, it must be a Nash equilibrium, and it must be the choice of rational players, provided there is common knowledge of rationality.

There is no plausible set of epistemic conditions that imply the common knowledge of rationality. This perhaps explains the many non-obvious, indeed perplexing, arguments surrounding the iterated elimination of strongly dominated strategies, some of which are presented and analyzed below.

4.1 Epistemic Games

The Nash equilibrium criterion (§2.4) does not refer to the knowledge or beliefs of players. If players are Bayesian rational (§1.5), however, they then have beliefs concerning the behavior of the other players, and they maximize their expected utility by choosing best responses, given these beliefs. Thus, to investigate the implications of Bayesian rationality, we must incorporate beliefs into the description of the game.

An *epistemic game* \mathcal{G} consists of a normal form game with players $i = 1, \dots, n$ and a finite pure strategy set S_i for each player i , so $S = \prod_{i=1}^n S_i$ is the set of pure strategy profiles for \mathcal{G} , with payoffs $\pi_i : S \rightarrow \mathbf{R}$. In addition, \mathcal{G} includes a set of possible states Ω of the game, a knowledge partition \mathcal{P}_i

of Ω for each player i , and a subjective prior (§1.5) $p_i(\cdot; \omega)$ over Ω that is a function of the current state ω . This subjective prior, $p_i(\cdot; \omega)$, is precisely the player's beliefs concerning the state of the game, including the choices of the other players. A state ω specifies, possibly among other aspects of the game, the strategy profile s used in the game. We write this $s = \mathbf{s}(\omega)$. Similarly, we write $s_i = \mathbf{s}_i(\omega)$ and $s_{-i} = \mathbf{s}_{-i}(\omega)$.

Recall from §1.6 that a *partition* of a set X is a set of mutually disjoint subsets of X whose union is X . We write the cell of the partition \mathcal{P}_i containing state ω as $\mathbf{P}_i\omega$, and we interpret $\mathbf{P}_i\omega \in \mathcal{P}_i$ as the set of states that i considers possible (i.e., among which i cannot distinguish) when the actual state is ω . Because player i cannot distinguish among states in $\mathbf{P}_i\omega$, his subjective prior must satisfy $p_i(\cdot; \omega) = p_i(\cdot; \omega')$ for all $\omega' \in \mathbf{P}_i\omega$. Finally, if $\psi: \Omega \rightarrow \{\text{true}, \text{false}\}$, we write $[\psi] = \{\omega \in \Omega \mid \psi(\omega) = \text{true}\}$.

The possibility operator \mathbf{P}_i has the following two properties: for all $\omega, \omega' \in \Omega$,

$$\begin{aligned} \text{(P1)} \quad & \omega \in \mathbf{P}_i\omega \\ \text{(P2)} \quad & \omega' \in \mathbf{P}_i\omega \Rightarrow \mathbf{P}_i\omega' = \mathbf{P}_i\omega \end{aligned}$$

P1 says that the current state is always possible, and P2 says that if ω' is possible at ω , then any state possible at ω is possible at ω' (i.e., $\mathbf{P}_i\omega \subseteq \mathbf{P}_i\omega'$) and because the condition is symmetrical in ω and ω' , if ω' is possible at ω , then ω is possible at ω' .

We call a set $E \subseteq \Omega$ an *event*, and we say that player i *knows* the event E at state ω if $\mathbf{P}_i\omega \subseteq E$; i.e., $\omega' \in E$ for all states ω' that i considers possible at ω . We write $\mathbf{K}_i E$ for the event that i knows E .

Given a possibility operator \mathbf{P}_i , we define the *knowledge operator* \mathbf{K}_i by

$$\mathbf{K}_i E = \{\omega \mid \mathbf{P}_i\omega \subseteq E\}.$$

The most important property of the knowledge operator is $\mathbf{K}_i E \subseteq E$; i.e., if an agent knows an event E in state ω (i.e., $\omega \in \mathbf{K}_i E$), then E is true in state ω (i.e., $\omega \in E$). This follows directly from P1.

We can recover the possibility operator $\mathbf{P}_i\omega$ for an individual from his knowledge operator \mathbf{K}_i , because

$$\mathbf{P}_i\omega = \bigcap \{E \mid \omega \in \mathbf{K}_i E\}. \quad (4.1)$$

To verify this equation, note that if $\omega \in \mathbf{K}_i E$, then $\mathbf{P}_i\omega \subseteq E$, so the left hand side of (4.1) is contained in the right hand side. Moreover, if ω' is

not in the right hand side, then $\omega' \notin E$ for some E with $\omega \in \mathbf{K}_i E$, so $\mathbf{P}_i \omega \subseteq E$, so $\omega' \notin \mathbf{P}_i \omega$. Thus the right hand side of (4.1) is contained in the left.

To visualize a partition \mathcal{P} of the universe into knowledge cells $\mathbf{P}_i \omega$, think of the universe Ω as a large corn field consisting of a rectangular array of equally spaced stalks. A fence surrounds the whole corn field, and fences running north/south and east/west between the rows of corn divide the field into plots, each completely fenced in. States ω are stalks of corn. Each plot is a cell $\mathbf{P}_i \omega$ of the partition, and for any event (set of corn stalks) E , $\mathbf{K}_i E$ is the set of plots completely contained in E (Collins 1997).

For example, suppose $\Omega = S = \prod_{i=1}^n S_i$, where S_i is the set of pure strategies of player i in a game \mathcal{G} . Then, one event is $P_{3t} = \{s = (s_1, \dots, s_n) \in \Omega \mid s_3 = t \in S_3\}$. This is the event that player 3 uses pure strategy t . More generally, if \mathcal{P}_i is i 's knowledge partition, and if i knows his own choice of pure strategy, but not that of the other players, each $P \in \mathcal{P}_i$ has the form $P_{it} = \{s = (t, s_{-i}) \in S \mid t \in S_i, s_{-i} \in S_{-i}\}$. Note that if $t, t' \in S_i$, then $t \neq t' \Rightarrow P_{it} \cap P_{it'} = \emptyset$, and $\cup_{t \in S_i} P_{it} = \Omega$, so \mathcal{P}_i is indeed a partition of Ω .

If \mathbf{P}_i is a possibility operator for i , the sets $\{\mathbf{P}_i \omega \mid \omega \in \Omega\}$ form a partition \mathcal{P} of Ω . Conversely, any partition \mathcal{P} of Ω gives rise to a possibility operator \mathbf{P}_i , two states ω and ω' being in the same cell iff $\omega' \in \mathbf{P}_i \omega$. Thus, a knowledge structure can be characterized by its knowledge operator \mathbf{K}_i , its possibility operator \mathbf{P}_i , or by its partition structure \mathcal{P} .

To interpret the knowledge structure, think of an event as a set of possible worlds in which some proposition is true. For instance, suppose E is the event "it is raining somewhere in Paris," and let ω be a state in which Alice is walking through the Jardin de Luxembourg where it is raining. Because the Jardin de Luxembourg is in Paris, $\omega \in E$. Indeed, in every state $\omega' \in \mathbf{P}_A \omega$ that Alice believes is possible, it is raining in Paris, so $\mathbf{P}_A \omega \subseteq E$; i.e., Alice knows that it is raining in Paris. Note that $\mathbf{P}_A \omega \neq E$, because, for instance, there is a possible world $\omega' \in E$ in which it is raining in Montmartre but not in the Jardin de Luxembourg. Then, $\omega' \notin \mathbf{P}_A \omega$, but $\omega \in E$.

Since each state ω in epistemic game \mathcal{G} specifies the players' pure strategy choices $\mathbf{s}(\omega) = (\mathbf{s}_1(\omega), \dots, \mathbf{s}_n(\omega)) \in S$, the players' subjective priors must specify their beliefs $\phi_1^\omega, \dots, \phi_n^\omega$ concerning the choices of the other players. We have $\phi_i^\omega \in \Delta S_{-i}$, which allows i to assume other player's choices are correlated. This is because, while the other players choose independently,

they may have communalities in beliefs that lead them independently to choose correlated strategies.

We call ϕ_i^ω i 's *conjecture* concerning the behavior of the other players at ω . Player i 's conjecture is derived from i 's subjective prior by noting that $[s_{-i}] =_{\text{def}} [\mathbf{s}_{-i}(\omega) = s_{-i}]$ is an event, so we define $\phi_i^\omega(s_{-i}) = p_i([s_{-i}]; \omega)$, where $[s_{-i}] \subset \Omega$ is the event that the other players choose strategy profile s_{-i} . Thus, at state ω , each player i takes the action $\mathbf{s}_i(\omega) \in S_i$ and has the subjective prior probability distribution ϕ_i^ω over S_{-i} . A player i is deemed *Bayesian rational* at ω if $\mathbf{s}_i(\omega)$ maximizes $\pi_i(s_i, \phi_i^\omega)$, where

$$\pi_i(s_i, \phi_i^\omega) =_{\text{def}} \sum_{s_{-i} \in S_{-i}} \phi_i^\omega(s_{-i}) \pi_i(s_i, s_{-i}). \quad (4.2)$$

In other words, player i is Bayesian rational in epistemic game \mathcal{G} if his pure strategy choice $\mathbf{s}_i(\omega) \in S_i$ for every state $\omega \in \Omega$, satisfies

$$\pi_i(\mathbf{s}_i(\omega), \phi_i^\omega) \geq \pi_i(s_i, \phi_i^\omega) \quad \text{for } s_i \in S_i. \quad (4.3)$$

We take the above to be the standard description of an epistemic game, so we assume without comment that if \mathcal{G} is an epistemic game, then the players are $i = 1, \dots, n$, the state space is Ω , the strategy profile at ω is $\mathbf{s}(\omega)$, the conjectures are ϕ_i^ω , i 's subjective prior at ω is $p_i(\cdot|\omega)$, and so on.

4.2 A Simple Epistemic Game

Suppose Alice and Bob each choose heads (h) or tails (t), neither observing the other's choice. We can write the universe as $\Omega = \{\text{hh}, \text{ht}, \text{th}, \text{tt}\}$, where xy means Alice chooses x and Bob chooses y . Alice's knowledge partition is then $\mathcal{P}_A = \{\{\text{hh}, \text{ht}\}, \{\text{th}, \text{tt}\}\}$ and Bob's knowledge partition is $\mathcal{P}_B = \{\{\text{hh}, \text{th}\}, \{\text{ht}, \text{tt}\}\}$. Alice's possibility operator \mathbf{P}_A satisfies $\mathbf{P}_A \text{hh} = \mathbf{P}_A \text{ht} = \{\text{hh}, \text{ht}\}$ and $\mathbf{P}_A \text{th} = \mathbf{P}_A \text{tt} = \{\text{th}, \text{tt}\}$, whereas Bob's possibility operator \mathbf{P}_B satisfies $\mathbf{P}_B \text{hh} = \mathbf{P}_B \text{th} = \{\text{hh}, \text{th}\}$ and $\mathbf{P}_B \text{ht} = \mathbf{P}_B \text{tt} = \{\text{ht}, \text{tt}\}$.

In this case, the event "Alice chooses h" is $E_A^h = \{\text{hh}, \text{ht}\}$, and because $\mathbf{P}_A \text{hh}, \mathbf{P}_A \text{ht} \subset E$, Alice knows E_A^h whenever E_A^h occurs (i.e., $E_A^h = \mathbf{K}_i E_A^h$). The event E_B^h expressing "Bob's chooses h" is $E_B^h = \{\text{hh}, \text{th}\}$, and Alice does not know E_B^h , because at th , Alice believes tt is possible, but $\text{tt} \notin E_B^h$.

4.3 An Epistemic Battle of the Sexes

Consider the Battle of the Sexes (§2.8), depicted to the right. Suppose there are four types of Violetta, V_1, \dots, V_4 , and four types of Alfredo, A_1, \dots, A_4 . Violetta V_1 plays $t_1 = o$ and conjectures that Alfredo chooses o . Violetta V_2 plays $t_2 = g$ and conjectures that Alfredo chooses g . Violetta V_3 plays $t_3 = g$ and conjectures that Alfredo plays his mixed strategy best response. Finally, Violetta V_4 plays $t_4 = o$ and conjectures that Alfredo plays his mixed strategy best response. Correspondingly, Alfredo A_1 plays $s_1 = o$ and conjectures that Violetta chooses o . Alfredo A_2 plays $s_2 = g$ and conjectures that Violetta plays g . Alfredo A_3 plays $s_3 = g$ and conjectures that Violetta plays her mixed strategy best response. Finally, Alfredo A_4 plays $s_4 = o$ and conjectures that Violetta plays her mixed strategy best response.

		Violetta	
		g	o
Alfredo	g	2,1	0,0
	o	0,0	1,2

A state of the game is $\omega_{ij} = (A_i, V_j, s_i, t_j)$, where $i, j = 1, \dots, 4$. We write $\omega_{ij}^A = A_i$, $\omega_{ij}^V = V_j$, $\omega_{ij}^s = s_i$, $\omega_{ij}^t = t_j$.

Define $E_i^A = \{\omega_{ij} \in \Omega | \omega_{ij}^A = A_i\}$ and $E_j^V = \{\omega_{ij} \in \Omega | \omega_{ij}^V = V_j\}$. Then, E_i^A is the event that Alfredo's type is A_i , and E_j^V is the event that Violetta's type is V_j . Since each type is associated with a given pure strategy, Alfredo's knowledge partition is $\{E_i^A, i = 1, \dots, 4\}$ and Violetta's knowledge partition is $\{E_j^V, j = 1, \dots, 4\}$.

Note that both players are Bayesian rational at each state of the game, because each strategy choice is a best response to the player's conjecture. Also, a Nash equilibrium occurs at ω_{11} , ω_{22} , ω_{33} , and ω_{44} , although at only the first two of these are the players' conjectures correct. Of course, there is no mixed strategy Nash equilibrium, because each player chooses a pure strategy in each state. However, if we define a Nash equilibrium *in conjectures* at a state as a situation in which each player's conjecture is a best response to the other player's conjecture, then ω_{ii} is a Nash equilibrium in conjectures for $i = 1, \dots, 4$, and ω_{34} and ω_{43} are also equilibria in conjectures. Note that in this case, if Alfredo and Violetta have common priors and mutual knowledge of rationality, their choices form a Nash equilibrium in conjectures. We will generalize this in Theorem 8.2.

4.4 Dominated and Iteratedly Dominated Strategies

We say $s'_i \in S_i$ is *strongly dominated* by $s_i \in S_i$ if, for every $\sigma_{-i} \in \Delta^* S_{-i}$, $\pi_i(s_i, \sigma_{-i}) > \pi_i(s'_i, \sigma_{-i})$. We say s'_i is *weakly dominated* by s_i if for every $\sigma_{-i} \in \Delta^* S_{-i}$, $\pi_i(s_i, \sigma_{-i}) \geq \pi_i(s'_i, \sigma_{-i})$, and for at least one choice of σ_{-i} the inequality is strict. A strategy may fail to be strongly dominated by any pure strategy, but may nevertheless be strongly dominated by a mixed strategy (§4.11).

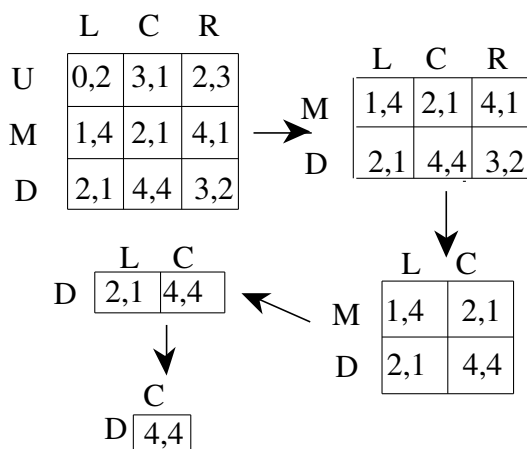


Figure 4.1. The Iterated Elimination of Strongly Dominated Strategies

Having eliminated dominated strategies for each player, it often turns out that a pure strategy that was not dominated at the outset is now dominated. Thus, we can undertake a second round of eliminating dominated strategies. Indeed, this can be repeated until no remaining pure strategy can be eliminated in this manner. In a finite game, this will occur after a finite number of rounds, and will always leave at least one pure strategy remaining for each player. If strongly (resp. weakly) dominated strategies are eliminated, we call this the *iterated elimination of strongly (resp. weakly) dominated strategies*. We call a pure strategy eliminated by this procedure an *iteratedly dominated strategy*.

Figure 4.1 illustrates the iterated elimination of strongly dominated strategies. First, U is strongly dominated by D for player 1. Second, R is strongly dominated by $0.5L + 0.5C$ for player 2 (note that a pure strategy in this case is not dominated by any other pure strategy, but is strongly dominated by a mixed strategy). Third, M is strongly dominated by D, and finally, L is

strongly dominated by C. Note that $\{D,C\}$ is indeed the unique Nash equilibrium of the game.

4.5 Eliminating Weakly Dominated Strategies

This example, due to Rubinstein (1991), starts with the Battle of the Sexes game \mathcal{G} (§2.8), where if players choose gg, Alfredo gets 3 and Violetta gets 1, if they choose oo, Alfredo gets 1 and Violetta gets 3, and if they choose og or go, both get nothing. All three subgames of \mathcal{G} are subgame perfect, because there are no proper subgames. Now, suppose Alfredo says to Violetta before they make their choices, “I have the option of throwing away 1 before I choose, if I so desire. Now the new game \mathcal{G}^+ is shown in Figure 4.2.

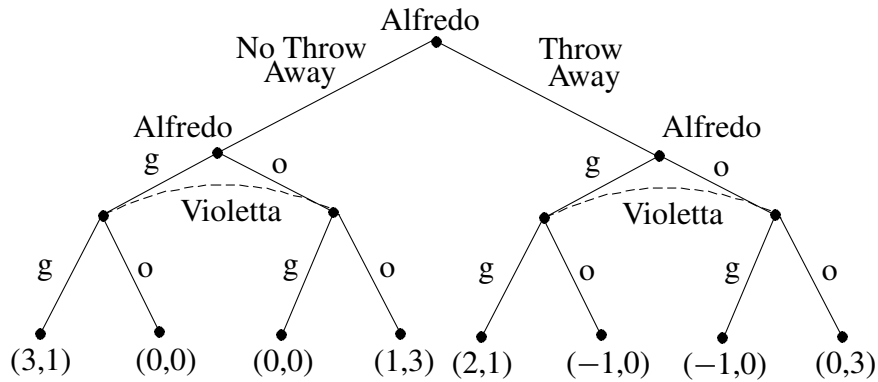


Figure 4.2. Battle of Sexes with Money Burning

This game has many Nash equilibria, all of which are subgame perfect, since there are not proper subgames to \mathcal{G}^+ . Suppose we *ignore* the extensive form structure of the game, and apply the iterated elimination of weakly dominated strategies to the normal form of this game. The normal form is shown in Figure 4.3, where nx means “don’t burn, choose x”, bx means “burn money (throw away 1) and choose x”, gg means “choose g”, oo means “choose o”, go means “choose g if Alfredo does not burn and o if Alfredo burns,” and og means “choose o if Alfredo does not burn and g if Alfredo burns.”

Now, bo is weakly dominated by ng, then oo is weakly dominated by og, then go is weakly dominated by gg, then no is weakly dominated by bg, then og is weakly dominated by gg, and then bg is weakly dominated by

	gg	go	og	oo
ng	3,1	3,1	0,0	0,0
no	0,0	0,0	1,3	1,3
bg	2,1	-1,0	2,1	-1,0
bo	-1,0	0,3	-1,0	0,3

Figure 4.3. Normal Form of Battle of Sexes with Money Burning

ng, leaving only the Nash equilibrium (ng,gg). Thus we have found that a purely hypothetical possibility that Alfredo might “burn money,” although he never does so, allows him to enjoy the high-payoff Nash equilibrium in which he earns 3 and Violetta earns 1.

4.6 Rationalizable Strategies

Suppose \mathcal{G} is an epistemic game. We denote the set of mixed strategies with support in S as $\Delta^* S = \prod_{i=1}^n \Delta S_i$, where ΔS_i is the set of mixed strategies for player i . We denote the mixed strategy profiles of all $j \neq i$ by $\Delta^* S_{-i}$.

In §1.5, we found that an agent whose choices satisfy the Savage axioms behaves as if maximizing a preference function subject to a subjective prior over the states of nature. We tailor this definition to epistemic game theory by saying player i is rational at state ω if his pure strategy $s_i(\omega)$ is a best response to his conjecture ϕ_i^ω of the other players’ strategies at ω (equation 4.3). Since a strongly dominated strategy can never be a best response, it follows that a rational player will never use a strongly dominated strategy. Moreover, if i knows that j is rational, and hence will never use a strongly dominated strategy, then i can eliminate pure strategies in S_i that are best responses only to strategies in $\Delta^* S_{-i}$ that do not use strongly dominated pure strategies in S_{-i} . Moreover, if i knows that j knows that k is rational, then i knows that j will eliminate pure strategies that are best responses to k ’s strongly dominated strategies, and hence i can eliminate pure strategies that are best replies only to j ’s eliminated strategies. And so on. Pure strategies that survive this back-and-forth iterated elimination of pure strategies are called *rationalizable* (Bernheim 1984, Pearce 1984).

One elegant formal characterization of rationalizable strategies is in terms of *best response sets*. In epistemic game \mathcal{G} , we say a set $X = \prod_{i=1}^n X_i$,

where each $X_i \subseteq S_i$, is a best response set if, for each i and each $x_i \in X_i$, i has a conjecture $\phi_i \in \Delta X_{-i}$ such that x_i is a best response to ϕ_{-i} , as defined by (4.3). It is clear that the union of two best response sets is also a best response set, so the union of all best response sets is a maximal best response set. We define a strategy to be *rationalizable* if it is a member of this maximal best response set.

Note that the pure strategies for each player used with positive probability in a Nash equilibrium form a best response set, in which each player conjectures the actual mixed strategy choice of the other players. Therefore, any pure strategy used with positive probability in a Nash equilibrium is rationalizable. In a game with a completely mixed Nash equilibrium (§2.3), it follows that all strategies are rationalizable.

This definition of rationalizability is not constructive; i.e., knowing the definition does not tell us how to find the set that satisfies it. The following construction leads to the same set of rationalizable strategies. Let $S_i^0 = S_i$ for all i . Having defined S_i^k for all i and for $k = 0, \dots, r-1$, we define S_i^r to be the set of pure strategies in S_i^{r-1} that are best responses to some conjecture $\phi_i \in \Delta S_{-i}^{r-1}$. Since $S_i^r \subseteq S_i^{r-1}$ for each i and there are only finite number of pure strategies, there is some $r > 0$ such that $S_i^r = S_i^{r-1}$, and clearly for any $l > 0$, we then have $S_i^r = S_i^{r+l}$. We define i 's rationalizable strategies as S_i^r .

These constructions refer only obliquely to the game's epistemic conditions, and in particular to the common knowledge of rationality (CKR) on which the rationalizability criterion depends. CKR obtains when each player is rational, each knows the others are rational, each knows the others know the others are rational, and so on. There is a third construction of rationalizability that makes its relationship to common knowledge of rationality more transparent.

Let s_1, \dots, s_n be the strategy profile chosen when ϕ_1, \dots, ϕ_n are the players' conjectures. The rationality of player i requires that s_i maximize i 's expected payoff, given ϕ_i . Moreover, because i knows that j is rational, he knows that s_j is a best response, given some probability distribution over S_{-j} —namely, s_j is a best response to ϕ_j . We say ϕ_i is *first-order consistent* if ϕ_i places positive probability only on pure strategies of j that have the property of being best responses, given some probability distribution over S_{-j} . By the same reasoning, if i places positive probability on the pair s_j, s_k , because i knows that j knows that k is rational, i knows that j 's conjecture is first-order consistent, and hence i places positive prob-

ability only on pairs s_j, s_k where j is first-order consistent and j places positive probability on s_k . When this is the case, we say that i 's conjecture is *second-order consistent*. Clearly, we can define consistency of order r for all positive integers r , and a conjecture which is r -consistent for all r is simply called *consistent*. We say s_1, \dots, s_n is rationalizable if there is some consistent set of conjectures ϕ_1, \dots, ϕ_n that places positive probability on s_1, \dots, s_n .

I leave it to the reader to prove that these three constructions define the same set of rationalizable strategies.

4.7 Eliminating Strongly Dominated Strategies

Consider the constructive approach to rationalizability developed in §4.6. It is clear that a strongly dominated strategy will be eliminated in the first round of the rationalizability construction if and only if it is eliminated in the first round of the iterated elimination of strongly dominated strategies. This observation can be extended to each successive stage in the construction of rationalizable strategies, which shows that all strategies that survive the iterated elimination of strongly dominated strategies are rationalizable. Are there other strategies that are rationalizable? The answer is that the strongly dominated strategies exhaust the rationalizable strategies, given our assumption that players can have correlated conjectures. For details, see Bernheim (1984) or Pearce (1984).

4.8 Common Knowledge of Rationality

We will now define CKR formally. Let \mathcal{G} be an epistemic game. For conjecture $\phi_i \in \Delta S_{-i}$, define $\text{argmax}_i(\phi_i) = \{s_i \in S_i \mid s_i \text{ maximizes } \pi_i(s'_i, \phi_i)\}$; i.e., $\text{argmax}_i(\phi_i)$ is the set of i 's best responses to the conjecture ϕ_i . Let $B_i(X_{-i})$ be the set of pure strategies of player i that are best responses to some mixed strategy profile $\sigma_{-i} \in X_{-i} \subseteq S_{-i}$; i.e., $B_i(X_{-i}) = \{s_i \in S_i \mid (\exists \phi_i \in \Delta^* X_{-i}) s_i \in \text{argmax}_i(\phi_i)\}$. We abbreviate $\phi([s_j(\omega) = s_j]) > 0$ as $\phi(s_j) > 0$, and $\phi([s_{-i}(\omega) = s_{-i}]) > 0$ as $\phi(s_{-i}) > 0$. We define

$$K_i^1 = [(\forall j \neq i) \phi_i^\omega(s_j) > 0 \Rightarrow s_j \in B_j(S_{-j})]. \quad (4.4)$$

K_i^1 is thus the event that i conjectures that a player j chooses s_j , only if s_j is a best response for j . In other words, K_i^1 is the event that i knows the other players are rational.

Suppose we have defined K_i^k for $k = 1, \dots, r-1$. We define

$$K_i^r = K_i^{r-1} \cap [(\forall j \neq i) \phi_i^\omega(s_j) > 0 \Rightarrow s_j \in B_j(K_j^{r-1})].$$

Thus, K_i^2 is the event that i knows that every player knows that every player is rational. Similarly, K_i^r is the event that i knows that every chain of r recursive “ j knows that k .” We define $K^r = \bigcap_i K_i^r$, and if $\omega \in K^r$, we say there is *mutual knowledge of degree r* . Finally, we define the event CKR as

$$K^\infty = \bigcap_{r \geq 1} K^r.$$

Note that in an epistemic game, CKR cannot simply be *assumed*, and is not a property of the players or of the informational structure of the game. This is because CKR generally holds only in certain states, and fails in other states. For example, in Chapter 5, we prove Aumann’s famous theorem that in a generic extensive form game of perfect information, where distinct states are associated with distinct choice nodes, CKR holds only at nodes on the backward induction path (§5.11). The confusion surrounding CKR generally flows from attempting to abstract from the epistemic apparatus erected to define CKR, and then to consider CKR to be some “higher form” of rationality that, when violated, impugns Bayesian rationality itself. There is no justification for such reasoning. There is nothing “irrational” about the failure of CKR. Nor is CKR some sort of “ideal” rationality that “boundedly rational” agents lamentably fail to attain.

4.9 Rationalizability and Common Knowledge of Rationality

We will use the following characterization of rationalizability (§4.6). Let $S_i^0 = S_i$ for all i , and define $S^0 = \prod_{i=1}^n S_i^0$ and $S_{-i}^0 = \prod_{j \neq i} S_j^0$. Having defined S^k and S_{-i}^k for all i and for $k = 0, \dots, r-1$, we define $S_i^r = B_i(S_{-i}^{r-1})$. Then, $S^r = \prod_{i=1}^n S_i^r$ and $S_{-i}^r = \prod_{j \neq i} S_j^r$. We call S^r the set of pure strategies that survive r iterations of the elimination of unrationalizable strategies. Since $S_i^r \subseteq S_i^{r-1}$ for each i and there are only finite number of pure strategies, there is some $r > 0$ such that $S_i^r = S_i^{r-1}$, and for any $l > 0$, we then have $S_i^r = S_i^{r+l}$. We define i ’s rationalizable strategies as S_i^r .

THEOREM 4.1 *For all players i and $r \geq 1$, if $\omega \in K_i^r$ and $\phi_i^\omega(s_{-i}) > 0$, then $s_{-i} \in S_{-i}^r$.*

This implies that if there is mutual knowledge of degree r at ω , and i 's conjecture at ω places strictly positive weight on s_{-i} , then s_{-i} survives r iterations of the elimination of unrationalizable strategies.

To prove this theorem, let $\omega \in K^1$ and suppose $\phi_i^\omega(s_j) > 0$. Then, $s_j \in B_j(S_{-j})$, and therefore $s_j \in S_j^1$, using the conjecture that maximizes s_j in $B_j(S_{-i})$. Since this is true for all $j \neq i$, $\phi^\omega(s_{-i}) > 0$ implies $s_{-i} \in S_{-i}^1$.

Now suppose we have proved the theorem for $k = 1, \dots, r$ and let $\omega \in K_i^{r+1}$. Suppose $\phi_i^\omega(s_j) > 0$. We will show that $\omega \in S_j^{r+1}$. By the inductive hypothesis and the fact that $\omega \in K_i^{r+1} \subseteq K_i^r$, we have $s_j \in S_j^r$, so s_j is a best response to some $\phi_j \in S_{-j}^r$. But then $s_j \in S_j^{r+1}$ by construction. Since this is true for all $j \neq i$, if $\phi_i^\omega(s_{-i}) > 0$, then $s_{-i} \in S_{-i}^{r+1}$.

4.10 The Beauty Contest

In his overview of behavioral game theory Camerer (2003) summarizes a large body of evidence in the following way: “Nearly all people use one step of iterated dominance. . . However, at least 10% of players seem to use each of two to four levels of iterated dominance, and the median number of steps of iterated dominance is two.” (p. 202) Camerer’s observation would be unambiguous if the issue were decision theory, where a single agent faces a non-strategic environment. But, in strategic interaction, the situation is more complicated. In the games reported in Camerer (2003), players gain by using one more level of backward induction than the other players. Hence, players must assess not how many rounds of backward induction the others are capable of, but rather how many the other players believe that other players will use. There is obviously an infinite recursion here, with little hope that considerations of Bayesian rationality will guide one to an answer. All we can say is that a Bayesian rational player will maximize expected payoff using a subjective prior over the expected number of rounds over which his opponents backward inducts. The beauty contest game (Moulin 1986) is crafted to explore this issue.

In the beauty contest game, each of $n > 2$ players chooses a whole number between zero and 100. Suppose the average of these n numbers is k . Then, the players whose choice is closest to $2k/3$ share a prize equally. It is obviously strongly dominated to choose a number greater than $2/3 \times 100 \approx 67$, because such a strategy has payoff zero, whereas the mixed strategy playing zero to 67 with equal probability has a strictly positive payoff. Thus, one-round of eliminating strongly dominated strategies eliminates choices

above 67. A second round of eliminating strongly dominated strategies eliminates choices above $(2/3)^2 \times 100 \approx 44$. Continuing in this manner, we see that the only rationalizable strategy is to choose zero. But, this is a poor choice in real life. Nagel (1995) studied this game experimentally with various groups of size 14 to 16. The average number chosen was 35, which is between two and three rounds of iterated elimination of strongly dominated strategies. This again conforms to Camerer's generalization, but in this case, of course, people play the game *far* from the Nash equilibrium of the game.

4.11 The Traveler's Dilemma

Consider the following game G_n , known as the *Traveler's Dilemma* (Basu 1994). Two business executives pay bridge tolls while on a trip, but do not have receipts. Their superior tells each of them to report independently an integral number of dollars between 2 and n on their expense sheet. If they report the same number, each will receive this much back. If they report different numbers, they each get the smaller amount, plus the low reporter gets an additional \$2 (for being honest), and the high reporter loses \$2 (for trying to cheat).

	s_2	s_3	s_4	s_5
s_2	2, 2	4, 0	4, 0	4, 0
s_3	0, 4	3, 3	5, 1	5, 1
s_4	0, 4	1, 5	4, 4	6, 2
s_5	0, 4	1, 5	2, 6	5, 5

Figure 4.4. The Traveler's Dilemma

Let s_k be the strategy "report k ". Figure 4.4 illustrated the game G_5 . Note first that s_5 is only weakly dominated by s_4 , but a mixed strategy $\epsilon s_2 + (1 - \epsilon)s_4$ strongly dominates s_5 whenever $1/2 > \epsilon > 0$. When we eliminate s_5 for both players, s_3 only weakly dominates s_4 , but a mixed strategy $\epsilon s_2 + (1 - \epsilon)s_3$ strongly dominates s_4 for any $\epsilon > 0$. When we eliminate s_4 for both players, s_2 strongly dominates s_3 for both players. Hence (s_2, s_2) is the only strategy pair that survives the iterated elimination of strongly dominated strategies. It follows that s_2 is the only rationalizable strategy, and the only Nash equilibrium as well.

The following exercise asks you to show that for $n > 3$, s_n in the game G_n is strongly dominated by a mixed strategy of s_2, \dots, s_{n-1} .

- Show that for any $n > 4$, s_n is strongly dominated by a mixed strategy σ_{n-1} using only s_{n-1} and s_2 .
- Show that eliminating s_n in G_n gives rise to the game G_{n-1} .
- Use the above reasoning to show that for any $n > 2$, the iterated elimination of strongly dominated strategies leaves only s_2 which is thus the only rationalizable strategy, and hence also the only Nash equilibrium of G_n .

Suppose $n = 100$. It is not plausible to think that individuals would actually play 2,2, because by playing a number greater than, say, 92, they are assured of at least 90.

4.12 The Modified Traveler's Dilemma

One might think that the problem is that pure strategies are dominated by mixed strategies, and as we will argue in chapter 6, rational agents have no incentive play mixed strategies in one-shot games.

However, we can change the game a bit so that 2,2 is the only strategy profile that survives the iterated elimination of pure strategies strictly dominated by pure strategies. In Figure 4.5, I have added 1% of s_2 to s_4 and 2% of s_2 to s_3 , for both players.

	s_2	s_3	s_4	s_5
s_2	2.00, 2.00	4.00, 0.04	4.00, 0.02	4.00, 0.00
s_3	0.04, 4.00	3.08, 3.08	5.08, 1.04	5.08, 1.00
s_4	0.02, 4.00	1.04, 5.08	4.04, 4.04	6.04, 2.00
s_5	0.00, 4.00	1.00, 5.08	2.00, 6.04	5.00, 5.00

Figure 4.5. The Modified Traveler's Dilemma

It is easy to check that now s_4 strictly dominates s_5 for both players, and when s_5 is eliminated, s_3 strictly dominates s_4 for both players. When s_4 is eliminated, s_2 strictly dominates s_3 .

This method will extend to a Modified Traveler’s Dilemma of any size. To implement this, let

$$f(m, q) = \begin{cases} q - 2 & q < m \\ q & q = m \\ m + 2 & q > m \end{cases},$$

and define

$$\begin{aligned} \pi(2, q) &= f(2, q) \quad \text{for } q = 2, \dots, n \\ \pi(m, q) &= \sum_{k=3, l=2, \dots, n} f(m, q) + f(2, q) \frac{n - k}{4(n + 1)} \end{aligned}$$

It is easy to show that this Modified Traveler’s Dilemma is strictly dominance solvable and the only rationalizable strategy again has payoff 2,2. Yet, it is clear that for large n , rational players would likely choose a strategy with payoff near n . This shows that there is something fundamentally wrong with the rationalizability criterion. The culprit is the CKR, which is the only questionable assumption we made in defining rationalizability. It is not irrational to choose a high number in the Modified Traveler’s Dilemma, and indeed doing so is likely to lead to a high payoff compared to the game’s only rationalizable strategy. However, doing so is not compatible with the common knowledge of rationality.

4.13 Global Games

Suppose Alice and Bob can cooperate (C) and earn 4, but by defecting (D) either can earn x , no matter what the other player does. However, if a player cooperates and the other does not, the cooperator earns zero. Clearly, if $x > 4$, D is a strictly dominant strategy, and if $x < 0$, C is a strictly dominant strategy. If $0 < x < 4$, the players have a Pareto-optimal strategy C in which they earn 4, but there is a second Nash equilibrium in which both players play D and earn $x < 4$.

		Bob	
		D	C
Alice	D	x, x	$x, 0$
C	$0, x$	$4, 4$	

Suppose, however, that x is private information, each player receiving an imperfect signal $\xi_i = x + \hat{\epsilon}_i$ that is uniformly distributed on the interval $[x - \epsilon/2, x + \epsilon/2]$, where $\hat{\epsilon}_A$ is independently distributed from $\hat{\epsilon}_B$. We can then demonstrate the surprising result that, no matter how small the error ϵ is, the resulting game has a unique rationalizable strategy which is

to play C for $x < 2$ and D for $x > 2$. Note that this is very far from the Pareto-optimal strategy, no matter how small the error.

To see that this is the only Nash equilibrium, note that a player surely chooses C when $\xi < -\epsilon/2$, and D when $\xi > 4 + \epsilon/2$, so there is a smallest cutoff x^* such that, at least in a small interval around x^* , the player chooses D when $\xi < x^*$ and C when $\xi > x^*$. For a discussion of this and other details of the model, see Carlsson and van Damme (1993), who invented and analyzed this game, which they term a *global game*. By the symmetry of the problem, x^* must be a cutoff for both players. If Alice is at the cutoff, then with equal probability Bob is above or below the cutoff, so he plays D and C with equal probability. This means that the payoff for Alice playing D is x^* and for playing C is 2. Because these must be equal if Alice is to have cutoff x^* , it follows that $x^* = 2$. Thus, there is a unique cutoff, and hence a unique Nash equilibrium $x^* = 2$.

To prove that $x^* = 2$ is the unique rationalizable strategy, suppose Alice chooses cutoff x_A and Bob chooses x_B as a best response. Then when Bob receives the signal $\xi_B = x_B$, he knows Alice's signal is uniformly distributed on $[x_B - \epsilon, x_B + \epsilon]$. To see this, let $\hat{\epsilon}_i$ be player i 's signal error, which is uniformly distributed on $[-\epsilon/2, \epsilon/2]$. Then

$$\xi_B = x + \hat{\epsilon}_B = \xi_A - \hat{\epsilon}_A + \hat{\epsilon}_B.$$

Because $-\hat{\epsilon}_A + \hat{\epsilon}_B$ is the sum of two random variables distributed uniformly on $[-\epsilon/2, \epsilon/2]$, ξ_B must be uniformly distributed on $[-\epsilon, \epsilon]$. It follows that the probability that Alice's signal is less than x_A is $q \equiv (x_A - x_B + \epsilon)/(2\epsilon)$, provided this is between zero and one. Then, x_B is determined by equating the payoff to D and C for Bob, which gives $4q = x_B$. Solving for x_B , we find that

$$x_B = \frac{2(x_A + \epsilon)}{2 + \epsilon} = x_A - \frac{(x_A - 2)\epsilon}{2 + \epsilon}. \quad (4.5)$$

The largest candidate for Alice's cutoff is $x_A = 4$, in which case Bob will choose cutoff $f_1 \equiv 4 - 2\epsilon/(2 + \epsilon)$. This means that no cutoff for Bob that is greater than f_1 is a best response for Bob, and therefore no such cutoff is rationalizable. But then the same is true for Alice, so the highest possible cutoff is f_1 . Now, using (4.5) with $x_A = f_1$, we define $f_2 = 2(f_1 + \epsilon)/(2 + \epsilon)$, and we conclude that no cutoff greater than f_2 is rationalizable. We can repeat this process as often as we please, each iteration k defining $f_k = 2(f_{k-1} + \epsilon)/(2 + \epsilon)$. Because the $\{f_k\}$ are decreasing and positive, they must have a limit, and this must satisfy the equation $f = 2(f + \epsilon)/(2 + \epsilon)$,

which has solution $f = 2$. Another way to see this is to calculate f_k explicitly. We find that

$$f_k = 2 + 2 \left(\frac{2}{2 + \epsilon} \right)^k,$$

which converges to 2 as $k \rightarrow \infty$, no matter how small $\epsilon > 0$ may be. To deal with cutoffs below $x = 2$, note that (4.5) must hold in this case as well. The smallest possible cutoff is $x = 0$, so we define $g_1 = 2\epsilon/(2 + \epsilon)$, and $g_k = 2(g_{k-1} + \epsilon)/(2 + \epsilon)$ for $k > 1$. Then, similar reasoning shows that no cutoff below g_k is rationalizable for any $k \geq 1$. Moreover the $\{g_k\}$ are increasing and bounded above by 2. The limit is then given by solving $g = 2(g + \epsilon)/(2 + \epsilon)$, which gives $g = 2$. Explicitly, we have

$$g_k = 2 - 2 \left(\frac{2}{2 + \epsilon} \right)^k,$$

which converges to 2 as $k \rightarrow \infty$. This proves that the only rationalizable cutoff is $x^* = 2$.

When the signal error is large, the Nash equilibrium of this game is plausible, and experiments show that subjects often settle on behavior close to that predicted by the model. However, the model predicts a cutoff of 2 for all $\epsilon > 0$, and a jump to cutoff 4 for $\epsilon = 0$. This prediction is not verified experimentally. In fact, subjects tend to treat the public information and private information scenarios the same, and tend to implement the payoff-dominant outcome rather than the less efficient Nash equilibrium outcome (Heinemann, Nagel and Ockenfels 2004, Cabrales, Nagel and Armenter 2007).

4.14 CKR is a Condition, not a Premise

Rational agents go through some process of eliminating unrationalizable strategies. CKR implies that players continue eliminating as long as there is anything to eliminate. By contrast, as we have seen, the median number of steps of iterated dominance found in experiments is two, and few players use more than four (Camerer 2003). This evidence indicates that CKR does not hold in the games analyzed in this chapter. Yet, it is easy to construct games in which we would expect CKR to hold. For instance, consider the following ‘‘Benign’’ Centipede game. Alice and Bob take turns for 100

rounds. On each round $r < 100$, the player choosing can Cooperate, in which case we move to the next round, or the player can Quit, in which case each player has a payoff of $\$(1 - r/100)$. If neither Quits, at the end of the 100 rounds, each player gets \$10. Rather than play round by round, each player submits a strategy for the whole game, which is thus treated as a normal form game.

CKR on this game implies Alice and Bob will both choose 100, and they will each earn \$10. For, in the final round, because Bob is rational, he will choose Continue, to earn \$10 as opposed to $\$(1 - 100/100) = 0$ from Quitting. Since Alice knows that Bob is rational, she knows she will earn \$10 by Continuing, as opposed to \$0.01 by Quitting. Now, on round 98, Bob earns \$0.02 by Quitting, which is more than Continuing and having Alice Quit, in which case he earns \$0.01. However, Bob knows that Alice knows that Bob is rational, and Bob knows that Alice is rational. Hence, Bob knows that Alice will Continue, so he Continues on round 98. The argument is valid back to round 1, so CKR implies cooperation on each round.

There is little doubt but that real-life players will play the strategy dictated by CKR in this case, although they do not in the Beauty Contest, the Traveler's Dilemma, and many other such games. Yet, there are no epistemic differences in what the players know about each other in the Benign Centipede game as opposed to the other games discussed above. It follows that the notion that CKR is a *premise* concerning the knowledge agents have about one another is false. Rather, CKR is a *condition* that a strategy profile chosen by agents may or may not satisfy *ex post*. Depending upon the particular game played, and under identical epistemic conditions, CKR may or may not hold.

I have stressed that a central weakness of epistemic game theory is the manner in which it represents the commonality of knowledges across individuals. Bayesian rationality itself supplies no analytical principles that are useful in deducing that two individuals have mutual, much less common, knowledge of particular events. We shall later suggest epistemic principles that do give rise to common knowledge (e.g., Theorem 7.2), but these do not include common knowledge of rationality. To my knowledge, no one has ever proposed a set of epistemic conditions that jointly imply CKR. Pettit and Sugden (1989) conclude their critique of CKR by asserting that "the situation where the players are ascribed common knowledge of their rationality ought strictly to have no interest for game theory." (p. 182) Unless

and until someone comes up with a epistemic derivation of CKR that explains why it is plausible in the Benign Centipede game but not the Beauty Contest game, this advice of Pettit and Sugden deserves to be heeded.

For additional analysis of CKR as a premise, see §5.13.