Part I

Foundations

Part II

Bargaining

Part III

Auctions and Mechanism Design

Part IV

Matching

Part V

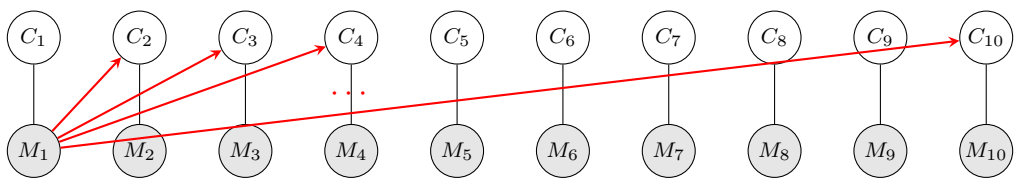
Information and Dynamic Games

Chapter 7

Common Knowledge

7.1 Motivating Example: The Corrupted Industry

Consider an industry with 10 firms $C_1, C_2, \ldots, C_{10}$. Each firm C_i has a manager M_i. The industry is thoroughly corrupted: every manager M_i is secretly selling their own company's private information to all other companies. In addition, each company is buying private information from the managers of other companies.



Every M_i sells to all C_j ($j \neq i$); only M_1's arrows shown.

There is one corporate governance rule: if a company discovers that its own manager is selling information to outsiders, it will fire the manager on the **Friday of that week**.

What does each firm know? Consider firm C_1. Since C_1 is buying information from the managers $M_2, M_3, \ldots, M_{10}$, it directly observes that those 9 managers are corrupt. However, C_1 **cannot observe whether its own manager** M_1 **is corrupt**—it does not know whether M_1 is selling to others. The same logic applies symmetrically to every firm. In summary:

- Each firm knows that *at least 9* managers are corrupt (the ones it buys from).
- No firm knows for certain whether its *own* manager is corrupt.

Now suppose that at the start of Week 1, a regulator makes a **public announcement**: *"At least one manager in this industry is corrupt."* This statement is heard by all 10 firms simultaneously, and everyone knows that everyone heard it.

What happens? Nothing happens in Week 1. Nothing happens in Week 2. In fact, nothing happens until **Week 10**, at which point all 10 managers are fired simultaneously.

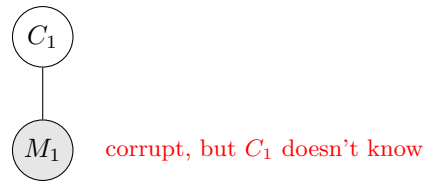
At first glance, this is deeply puzzling. Every firm already knew that at least 9 managers were corrupt—far more than the "at least one" in the announcement. How can such a

seemingly vacuous statement trigger any action at all, let alone a cascade that takes 10 weeks to unfold?

7.1.1 Iterated Reasoning

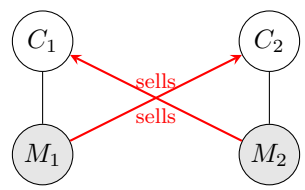
The answer lies in the chain of *iterated hypothetical reasoning* that the public announcement enables.

The $n = 1$ Case. Suppose there is only 1 firm with 1 corrupt manager.



Before the announcement, C_1 does not know its own manager is corrupt. After the announcement ("at least one is corrupt"), C_1 immediately deduces that M_1 must be the corrupt one, and fires him on Friday of **Week 1**.

The $n = 2$ Case. Suppose there are 2 firms, both with corrupt managers. Each firm observes that the other firm's manager is corrupt (by buying information from him), but does not know about its own.



Consider the reasoning from C_1's perspective after the announcement. C_1 observes that M_2 is corrupt, but entertains two hypotheses about its own manager:

- **Hypothesis A:** M_1 is not corrupt. Then only M_2 is corrupt. Under this hypothesis, after the public announcement, C_2 would find itself in the $n = 1$ situation: C_2 would see no other corrupt manager, learn from the announcement that at least one exists, and deduce that M_2 is the one. So C_2 would fire M_2 in **Week 1**.
- **Hypothesis B:** M_1 is also corrupt.

C_1 waits and watches. At the end of Week 1, M_2 is *not* fired. This rules out Hypothesis A: if M_1 had been clean, C_2 would have acted already. Therefore C_1 concludes that M_1 must also be corrupt, and fires M_1 in **Week 2**.

By perfectly symmetric reasoning, C_2 also watches whether C_1 fires M_1 in Week 1. When it does not happen, C_2 likewise infers that M_2 is corrupt. Both managers are fired simultaneously in **Week 2**.

The General Case: Induction. By induction on the number of corrupt managers n. Each firm sees $n - 1$ corrupt managers and entertains the hypothesis that its own manager is clean (so that only $n - 1$ managers are corrupt). Under that hypothesis, by the inductive assumption, the other $n - 1$ firms would have fired their managers by Week $n - 1$. When

© Rui Zhou 2026. Do not redistribute without permission.

this does not happen by the end of Week $n-1$, every firm deduces that the total must be n, meaning its own manager is also corrupt. All n managers are fired simultaneously in **Week** n.

For our example with $n = 10$: nothing happens for 9 weeks, and then all 10 managers are fired in Week 10.

7.1.2 Why Does a "Vacuous" Announcement Matter?

The content of the announcement ("at least one is corrupt") was already known to every firm—each firm could see at least 9 corrupt managers. What was *not* in place before the announcement was the full tower of *hypothetical reasoning about hypothetical reasoning*.

To see why, trace the chain from C_1's perspective. C_1 reasons: "I see 9 corrupt managers. But suppose M_1 is clean. Then C_2 would see only 8 corrupt managers. And C_2 might suppose M_2 is also clean. Then C_3 would see only 7..." Each level of nesting reduces the number of known-corrupt managers by one. After 9 levels of hypothetical reasoning, we reach a scenario where some firm sees *zero* other corrupt managers and does not know if any exist at all. Without the announcement, the induction has no base case in this deepest hypothetical—the chain of reasoning collapses before reaching its logical conclusion.

The public announcement plugs exactly this gap. By making "at least one is corrupt" publicly and simultaneously known to all, it provides the base case in every hypothetical scenario, no matter how deeply nested. This distinction—between knowledge that is shared but *finitely* nested, and knowledge that is shared *infinitely* deep—is precisely the distinction between *mutual knowledge* and *common knowledge*, which we now formalize.

7.2 Formal Definition

Definition 7.1: Mutual Knowledge and Common Knowledge

Let E be an event and let there be n agents.

- E is **mutual knowledge of order 1** if every agent knows E.
- E is **mutual knowledge of order** k if every agent knows that E is mutual knowledge of order $k-1$.
- E is **common knowledge** if E is mutual knowledge of order k for *every* $k \in \mathbb{N}$.

In other words, common knowledge means: everyone knows E, everyone knows that everyone knows E, everyone knows that everyone knows that everyone knows E, and so on *ad infinitum*.

In the corrupted industry example, the event $E =$ "at least one manager is corrupt" was, before the announcement, mutual knowledge of high order but **not** common knowledge. Each firm's chain of hypotheticals ("suppose mine is clean, then you see one fewer, then suppose yours is clean too...") bottoms out after finitely many levels at a scenario in which one firm sees zero corrupt managers and cannot determine whether E holds. Common knowledge requires that the chain never bottoms out—that at every depth of hypothetical

© Rui Zhou 2026. Do not redistribute without permission.

reasoning, every agent in the hypothetical scenario still knows E.

The public announcement achieves exactly this. Because every firm heard the announcement, and every firm knows every other firm heard it, and so on without bound, E becomes common knowledge. The base case of the induction ("a firm that sees zero corrupt managers and hears the announcement will fire its manager in Week 1") is now available at every depth, enabling the full chain of iterated reasoning to unfold.

7.3 Aumann's Agreement Theorem

A celebrated consequence of common knowledge is that two Bayesian agents who share a common prior cannot "agree to disagree" about the posterior of any event, provided the posteriors themselves are common knowledge.

Theorem 7.2: Aumann (1976)

Let two Bayesian agents have a common prior on a finite state space Ω and partitions $\mathcal{P}_1, \mathcal{P}_2$ representing what each agent observes. Let $E \subseteq \Omega$ be an event. If at state ω^* the values of both posteriors $\Pr_1(E \mid \mathcal{P}_1)$ and $\Pr_2(E \mid \mathcal{P}_2)$ are common knowledge, then they are equal.

Proof for Theorem

Let $q_1 = \Pr_1(E \mid \mathcal{P}_1(\omega^*))$ and $q_2 = \Pr_2(E \mid \mathcal{P}_2(\omega^*))$ be the posteriors. Common knowledge of q_1 at ω^* means that on every state in the meet-cell containing ω^*, agent 1's posterior is exactly q_1. Let $M = \mathcal{P}_1 \wedge \mathcal{P}_2$ be the meet (the join of the two partitions in the lattice of partitions, refined by both); the cell of M containing ω^* is the set of states agent 1's $\mathcal{P}_1$-cell could be, intersected with the corresponding $\mathcal{P}_2$-cells.

Within this meet-cell $M(\omega^*)$, agent 1's cells partition $M(\omega^*)$ into sub-cells $C_1, C_2, \ldots$ with $\Pr(E \mid C_k) = q_1$ on each. By the law of total probability,

$$\begin{aligned}\Pr(E \mid M(\omega^*)) &= \sum_k \Pr(C_k \mid M(\omega^*)) \cdot \Pr(E \mid C_k) \\ &= q_1 \cdot \sum_k \Pr(C_k \mid M(\omega^*)) \;=\; q_1.\end{aligned}$$

Symmetrically, $\Pr(E \mid M(\omega^*)) = q_2$. Hence $q_1 = q_2$. ∎

Remark (The Force of the Common-Prior Assumption).

The key non-triviality is that the agents need not have access to the same information; they only need to know the same prior and to commonly know each other's posterior. The argument iterates: each agent learns nothing about E beyond what is already encoded in the meet of partitions, because anything else would cause the posterior to update, breaking the common-knowledge assumption. Aumann's theorem has been used to argue against speculative trade between rational agents (the no-trade theorem of Milgrom-Stokey 1982): if both traders share a common prior and rationality is common knowledge, they cannot strictly benefit from trade, because the willingness to trade itself becomes informative.

© Rui Zhou 2026. Do not redistribute without permission.

7.4 The Coordinated Attack Problem

We now present a game-theoretic example that dramatically illustrates the difference between "almost common knowledge" and true common knowledge. This is based on Rubinstein's (1989) *Electronic Mail Game*.

7.4.1 Setup

Two divisions of an army (Player 1 and Player 2) must decide whether to **attack** (A) or **not attack** (N). The enemy is either **weak** (state G, the "good" state) or **strong** (state B, the "bad" state). The army wins if and only if *both* divisions attack *and* the enemy is weak. The prior probabilities are $\Pr(G) = \pi < \frac{1}{2}$ and $\Pr(B) = 1 - \pi$.

The payoff matrices are:

State G (prob π)

	A	N
A	$2,\ 2$	$-3,\ 0$
N	$0,\ -3$	$0,\ 0$

State B (prob $1 - \pi$)

	A	N
A	$-2,\ -2$	$-2,\ 0$
N	$0,\ -2$	$0,\ 0$

In state B, not attacking (N) is strictly dominant for both players. In state G, the game is a *coordination game*: (A, A) is a Nash equilibrium with payoff $(2, 2)$, but attacking alone is severely punished (-3). Crucially, if the state G were **common knowledge**, both (A, A) and (N, N) would be Nash equilibria—the game is a coordination game with two pure-strategy equilibria. The interesting question is whether the players can coordinate on the Pareto-superior equilibrium (A, A).

7.4.2 Information Structure

The communication protocol is as follows:

1. If the state is G, Nature sends a signal to Player 1 with certainty (no error at this stage). If the state is B, no signal is sent.

2. Upon receiving a signal, Player 1's system *automatically* forwards a message to Player 2. This transmission fails (the message is lost) with probability $\varepsilon > 0$, independently.

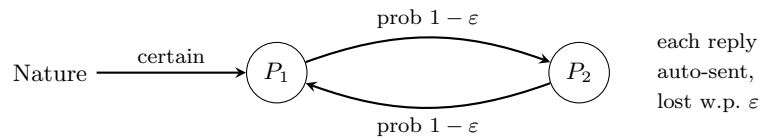
3. If Player 2 receives the message, Player 2's system automatically sends a confirmation back to Player 1, again lost with probability ε.

4. This back-and-forth continues indefinitely: each successfully received message triggers an automatic reply, each of which is independently lost with probability ε.

The message sending is *not* a strategic choice—it is automatic. The only strategic decision is whether to attack or not.

Let t_i denote the **number of messages received** by Player i. In state B, no messages are sent, so $t_1 = t_2 = 0$. In state G, the message chain proceeds as:

© Rui Zhou 2026. Do not redistribute without permission.



The chain stops at the first lost message. The possible outcomes given state G are:

Chain stops at	t_1	t_2
$P_1 \to P_2$ lost	1	0
$P_2 \to P_1$ lost	1	1
$P_1 \to P_2$ lost	2	1
$P_2 \to P_1$ lost	2	2
$\vdots$	$\vdots$	$\vdots$

Observe that t_1 and t_2 always differ by at most 1: given $t_1 = T \geq 1$, the value of t_2 is either $T-1$ or T.

Remark ("Almost" Common Knowledge).

When ε is small, the message chain is very likely to be long. With high probability, both players receive many confirmations, so each player is "almost sure" the state is G, "almost sure" the other knows it, "almost sure" the other knows the other knows it, and so on. The state G is "almost" common knowledge. One might expect that (A, A) should be sustainable as an equilibrium. As we will see, this intuition is spectacularly wrong.

7.4.3 The Unique Equilibrium

A strategy for Player i is a function $\sigma_i : \{0, 1, 2, \ldots\} \to \{A, N\}$ mapping the number of messages received to an action.

Proposition 7.3: Unique Equilibrium of the Coordinated Attack Game

There is a unique equilibrium: both players play N regardless of the number of messages received. That is, $\sigma_i(t_i) = N$ for all $t_i \geq 0$ and $i \in \{1, 2\}$.

Proof for Proposition.

By induction on the number of messages received.

Base case: $t_1 = 0$. If Player 1 receives no messages, then state B is certain (since Nature always signals Player 1 in state G). In state B, N is strictly dominant. So $\sigma_1(0) = N$.

Base case: $t_2 = 0$. If Player 2 receives no messages, there are two possibilities:

(i) State is B and no signal was ever sent (prior probability $1 - \pi$);

(ii) State is G, Player 1 received the signal, but the forward to Player 2 was lost (prior probability $\pi\varepsilon$).

© Rui Zhou 2026. Do not redistribute without permission.

By Bayes' rule:

$$\Pr(G \mid t_2 = 0) = \frac{\pi\varepsilon}{\pi\varepsilon + (1-\pi)}.$$

If Player 2 plays N, the payoff is 0. If Player 2 plays A, the payoff is at most (assuming Player 1 plays A whenever state is G):

$$\Pr(G \mid t_2 = 0) \times 2 + \Pr(B \mid t_2 = 0) \times (-2) = \frac{2\pi\varepsilon - 2(1-\pi)}{\pi\varepsilon + (1-\pi)} < 0,$$

since $\pi\varepsilon \ll 1 - \pi$ (recall $\pi < \frac{1}{2}$ and ε is small). So $\sigma_2(0) = N$.

Inductive step. Suppose that for all $t_1, t_2 < T$, both players play N: $\sigma_1(t_1) = N$ and $\sigma_2(t_2) = N$.

Consider Player 1 with $t_1 = T$ (the argument for Player 2 is symmetric). Given $t_1 = T$, Player 2's message count is either $t_2 = T - 1$ or $t_2 = T$. The conditional probability is:

$$p := \Pr(t_2 = T - 1 \mid t_1 = T) = \frac{\varepsilon}{\varepsilon + (1-\varepsilon)\varepsilon} = \frac{1}{2-\varepsilon} > \frac{1}{2}.$$

This is because, given $t_1 = T$: with probability ε the next message ($P_1 \to P_2$) is lost, yielding $t_2 = T - 1$; with probability $(1-\varepsilon)\varepsilon$ the message reaches P_2 (so $t_2 \geq T$) but the return is lost, yielding $t_2 = T$.

By the induction hypothesis, $\sigma_2(T-1) = N$. Even in the best case where Player 2 plays A when $t_2 = T$, if Player 1 plays A:

$$\text{Payoff to } P_1 \leq p \times (-3) + (1-p) \times 2 = \frac{-3 + 2(1-\varepsilon)}{2-\varepsilon} = \frac{-1-2\varepsilon}{2-\varepsilon} < 0.$$

The -3 arises because when $t_2 = T - 1$, Player 2 plays N (by induction), so Player 1 attacks alone. The 2 is the best-case payoff if Player 2 also attacks. Since $p > \frac{1}{2}$, the expected payoff from attacking is strictly negative. Therefore $\sigma_1(T) = N$.

By symmetric reasoning applied to Player 2, $\sigma_2(T) = N$. This completes the induction.

7.4.4 Interpretation

The result is striking. The following two scenarios yield completely different equilibrium outcomes:

Information structure	**Equilibrium outcome**
G is **common knowledge**	(A, A) is a NE
G is "almost" common knowledge	(N, N) is the *unique* NE

No matter how small $\varepsilon > 0$ is, and no matter how many thousands of confirmations each player has received, the *unique* equilibrium is for both to play N. The problem is not that the players lack confidence that G is true—after many messages, they are nearly certain. The problem is that each player is never quite sure the *other* player is sure that the other player is sure... that the state is G.

Remark (Common p-Belief).

© Rui Zhou 2026. Do not redistribute without permission.

The key quantitative insight is the conditional probability $p = \frac{1}{2-\varepsilon}$. Even when ε is tiny, p is close to $\frac{1}{2}$—far from 1. This means that no matter how many messages Player i has received, Player i always assigns probability at least $\frac{1}{2}$ to the event that the other player received one fewer message. The information structure "self-replicates" at every level: the uncertainty never shrinks as we go deeper in the epistemic hierarchy.

This connects to the concept of *common p-belief* (Monderer and Samet, 1989): an event is common p-belief if everyone believes it with probability at least p, everyone believes with probability at least p that everyone believes it with probability at least p, and so on. In the electronic mail game, the state G is common p-belief only for $p \leq \frac{1}{2-\varepsilon} \approx \frac{1}{2}$, which is far from the $p = 1$ required for common knowledge. Coordination requires common knowledge (or at least common p-belief with p sufficiently large); "almost common knowledge" is not enough.

Remark (Chapter Summary).

Common knowledge is the precise epistemic condition under which agents' inferences about each other's reasoning can iterate without bound: E is common knowledge if everyone knows E, everyone knows that everyone knows E, and so on *ad infinitum*. The chapter developed three pillars. *The fixed-point characterization*: an event is common knowledge in a state ω if and only if it is true at every state in the meet of the players' partitions containing ω. *Aumann's agreement theorem*: agents with a common prior who are commonly known to be Bayesian rational cannot "agree to disagree" about the posterior of any event—if their posteriors are common knowledge, they must coincide. *The fragility of common knowledge*: Rubinstein's electronic mail game shows that arbitrarily many rounds of mutual confirmation can fail to deliver common knowledge, and that small departures from common knowledge can flip the equilibrium discontinuously. The recurring theme: many central results in game theory (backward induction, Folk Theorem, signaling refinements) implicitly assume common knowledge of rationality, and the assumption matters more than its mathematical innocuousness suggests.

© Rui Zhou 2026. Do not redistribute without permission.

Part VI

Problem Sets and Solutions

Part VII

Exams and Solutions