Rationalizability

Game Theory · Module 2 — Nash Equilibrium

Node 2.5

Position in Knowledge Graph

System: Game Theory **Structural Domain:** Nash Equilibrium **Node:** 2.5

Rationalizability (Bernheim 1984, Pearce 1984) is a solution concept that sits between IESDS and Nash equilibrium. It asks: which strategies can be justified by some coherent belief that the opponent is also rational? The answer is the set of strategies surviving iterated deletion of *never-best-response* strategies — strategies that cannot be rationalized as a best response to any opponent belief. Rationalizability relaxes the Nash equilibrium requirement that beliefs be *correct*; it only requires that they be *rationalizable*. It is therefore weaker than Nash (every Nash strategy is rationalizable, but not conversely), and broader than IESDS (for 3+ players). It is the solution concept most closely aligned with pure common knowledge of rationality.

Formal Definition

Let $G = (N, \{S_i\}, \{u_i\})$ be a finite game with mixed extension.

A strategy $s_i \in S_i$ is a **best response to some belief** if there exists a probability distribution μ over S_{-i} such that $s_i \in \arg\max_{s_i' \in S_i} \sum_{s_{-i}} \mu(s_{-i}) u_i(s_i', s_{-i})$.

Define iteratively:

$$R_i^{(0)} = S_i, \qquad R_i^{(k+1)} = \{s_i \in R_i^{(k)} : s_i \text{ is a best response to some belief over } R_{-i}^{(k)}\}.$$

The **rationalizable strategies** of player i are

$$R_i^{\infty} = \bigcap_{k \geq 0} R_i^{(k)}.$$

Here $R_{-i}^{(k)}$ denotes the product $\prod_{j \neq i} R_j^{(k)}$, and beliefs over it are (in the standard Bernheim–Pearce formulation) required to be **independent** across opponents — each $j \neq i$ draws from their own $R_j^{(k)}$ independently. Some treatments allow correlated beliefs (Aumann), in which case rationalizability coincides with IESDS for all n.

A strategy s_i is **rationalizable** if $s_i \in R_i^{\infty}$.

Intuition

The idea is common knowledge of rationality, stripped of the Nash requirement that beliefs be correct.

A rational player cannot play a strategy that is never a best response: there would be no belief under which playing it is optimal. So rationality alone eliminates "never-best-response" strategies.

A rational player who knows the opponent is rational cannot justify a belief that assigns positive probability to the opponent's never-best-response strategies. So round-2 elimination throws out strategies that require irrationality of the opponent.

Iterating, at round k, a surviving strategy is one that can be justified by a chain of beliefs: "I believe the opponent believes I believe ... (k levels deep) ... that everyone is rational." Rationalizability is the fixed point of this recursion.

In Nash equilibrium, the player's belief about the opponent is *correct*: μ equals the opponent's actual σ^*_{-i}. In rationalizability, the player's belief is *coherent with common rationality* but need not be correct. This makes rationalizability strictly weaker.

Structural Role

Rationalizability bridges dominance and equilibrium:

- **IESDS $\Rightarrow$ rationalizability (for 2 players).** In two-player games, iterated strict dominance and rationalizability coincide (Pearce 1984).
- **For 3+ players:** rationalizability is weaker than IESDS under independent beliefs. Correlated-belief rationalizability coincides with IESDS for any n.
- **Rationalizability $\Rightarrow$ Nash existence.** If R_i^∞ is a singleton for each i, the unique profile is a Nash equilibrium.
- **Rationalizability $\subseteq$ Nash strategies.** Every strategy that is played with positive probability in some Nash equilibrium is rationalizable. The converse fails: some rationalizable strategies are not Nash.

Rationalizability also connects to interactive epistemology (Module 4): it is the solution concept for games with *common knowledge of rationality* and no equilibrium assumption. Nash equilibrium additionally requires common knowledge of the equilibrium profile itself, which is a strong assumption that rationalizability dispenses with.

Mathematical Development

Best-response sets

For a finite game, the set of strategies that are best responses to *some* mixed belief is well-defined:

$$\mathrm{BR}_i^{\mathrm{any}}(S_{-i}) = \bigcup_{\mu \in \Delta(S_{-i})} \mathrm{BR}_i(\mu).$$

The iteration $R_i^{(k+1)} = \mathrm{BR}_i^{\mathrm{any}}(R_{-i}^{(k)}) \cap R_i^{(k)}$ converges in finitely many steps for a finite game.

Pearce's Lemma

Lemma (Pearce, 1984). In a finite two-player game, a pure strategy s_i is a best response to some mixed opponent belief iff it is not strictly dominated (possibly by a mixed strategy).

This gives the equivalence between IESDS and rationalizability for two-player games: rationalizability-surviving = IESDS-surviving.

For more than two players, if beliefs are required to be independent (the standard definition), then there can be a strategy that is strictly dominated (even by a mixture) but *becomes* rationalizable because the dominating mixture requires correlation among opponents. The Bernheim–Pearce formulation therefore says rationalizability is slightly coarser than IESDS for $n \geq 3$ unless correlated beliefs are allowed.

Cournot revisited

Iteratively eliminating never-best-responses in Cournot duopoly converges to the unique Cournot–Nash equilibrium. The rationalizable set collapses to a single profile. In contrast, Bertrand competition with capacity constraints can have a large rationalizable set (many strategies are best responses to some belief).

Example: 3 strategies with no dominance but non-trivial rationalizability

	L	C	R
T	$(4,0)$	$(0,4)$	$(2,2)$
M	$(0,4)$	$(4,0)$	$(2,2)$
B	$(1,1)$	$(1,1)$	$(1,1)$

No pure strategy is strictly dominated. Is B rationalizable? Check: is there any belief under which B is a best response? Row's payoff from B is 1 regardless. Row's payoff from T is $4q_L + 0q_C + 2q_R \geq 2$ when $q_L \geq 1/2$ or $q_R \geq 1/2$. If $q_L = q_C = 0, q_R = 1$, Row gets 2 from T, 2 from M, 1 from B — B is not a best response. If $q_L = 0, q_C = 1$, Row gets 0, 4, 1 — B not best. It turns out B is never a best response: the mix $0.5T + 0.5M$ always yields ≥ 2, which beats 1. So B is dominated by $(1/2, 1/2, 0)$ — strictly dominated by a mixed strategy. Rationalizability (= IESDS here) deletes B. Remaining game has only T, M vs L, C, R.

Iterated best response as fictitious-play approximation

Rationalizability is the "limit" of infinitely iterated best response starting from the full strategy space. Fictitious play (Node 9.1) is a dynamic that realizes this iteratively, updating beliefs based on empirical opponent frequencies. In zero-sum and some potential games, fictitious play converges to a Nash equilibrium within the rationalizable set.

Worked Example

Computing rationalizable strategies

	L	C	R
T	$(5,2)$	$(1,3)$	$(0,0)$
M	$(3,4)$	$(0,0)$	$(2,1)$
B	$(0,0)$	$(4,1)$	$(3,2)$

Round 1. Any pure dominated strategy? Check Row: T = (5, 1, 0), M = (3, 0, 2), B = (0, 4, 3). No pure domination. Check mixed: $0.5M + 0.5B$ gives (1.5, 2, 2.5) — does this dominate T's (5, 1, 0)? No, T beats it in column L. No domination.

Column: L = (2, 4, 0), C = (3, 0, 1), R = (0, 1, 2). No domination.

Try: is there a strategy for either player that is not a best response to any belief? Row's T is a best response when Column's belief concentrates on L (Row gets 5). Row's M is a best response to L? $3q_L + 0q_C + 2q_R$ vs T's $5q_L + q_C + 0q_R$. For M to beat T, need $3q_L + 2q_R > 5q_L + q_C \Rightarrow -2q_L - q_C + 2q_R > 0$. With $q_L = 0, q_C = 0, q_R = 1$, this is $2 > 0$ ✓, but then we should also check if M beats B: 2 vs 3 — B is better. So M is not a best response to $(0, 0, 1)$. Try $(0, 0.5, 0.5)$: M gives 1, T gives 0.5, B gives 3.5 — M not best. Try $(0.5, 0, 0.5)$: M gives $1.5 + 1 = 2.5$, T gives 2.5, B gives 1.5. M ties with T, so is a best response. ✓

So M is rationalizable. By similar analysis, T and B are best responses to some beliefs, and all Column strategies are too. The rationalizable set is the entire game in this case — no elimination.

Nash equilibria lie in this set. Computing them is the next step (indifference equations).

Poker Anchor

Concept mapping

Formal object	Poker instantiation
Rationalizable strategy	A line that can be justified under *some* plausible read of opponent play
Never-best-response	A line that cannot be defended regardless of opponent tendencies — "just don't do this"
Iterated rationalizability	Successive rounds of "well, if they're rational, they wouldn't do X, and knowing that, I shouldn't do Y"
Rationalizable but not Nash	An exploit that works against some plausible opponent but is not GTO
Nash $\subset$ rationalizable	Every GTO action is a best response to GTO, hence rationalizable

Concrete situation

Multi-way pot on the river. You are in a 4-way pot on the river with a medium strength hand. UTG bets large, MP folds, CO calls, and action is on you. The decision space is "fold / call / raise."

Rationalizability analysis. Is "raise" a best response to *any* reasonable read? You would need to believe (a) UTG's betting range is weighted toward thin value and bluffs that will fold to a raise, (b) CO's calling range is weighted toward weak hands that will also fold. If you believe *both*, a raise is a best response. If you believe either is strong, a raise is catastrophic. So a raise is *rationalizable* if you can defend it with a coherent read — say, you've seen UTG bluff this exact spot before and CO is a known caller who folds to raises — but it is *not Nash* unless this belief is actually correct in the equilibrium population.

In Nash (GTO) analysis, the raise frequency in this spot is close to 0 — raising is rarely optimal in 4-way river pots because the ranges are too condensed. But the raise remains *rationalizable* because it is a best response to some belief about opponents' ranges.

Training implication. When a coach tells you "that play is defensible against this opponent type, but it's not GTO," they are saying the play is rationalizable but not Nash — it lies in $R_i^\infty \setminus \{\text{Nash supports}\}$. It is a play that can be justified, but only under specific beliefs, not under the common-rationality fixed point.

What this understanding buys you

Rationalizability is the right concept for exploitation. When you deviate from GTO to exploit a specific opponent, you are selecting a rationalizable strategy that is a best response to your read of them. As long as your read is coherent — i.e., the opponent's strategy is itself rationalizable given *their* reads of *you* — the play is defensible.

Don't confuse "rationalizable" with "good." Many rationalizable plays are losing plays if the belief they rely on is wrong. The discipline is to make the belief explicit, verify the belief is plausible given your information, and only then take the rationalizable-but-not-Nash line. The failure mode is taking a rationalizable line with a sloppy belief.

Nash is the unexploitable floor; rationalizability is the exploit ceiling. GTO is guaranteed not to lose to any opponent. The best rationalizable exploit is guaranteed to win the *most* against a specific opponent, *if* your read is correct. Live poker is the art of navigating between these two.

Cross-Links

- **1.3 Dominance and IESDS** — the upstream concept.

- **1.4 Best response** — the defining operator for rationalizability.
- **2.1 Nash equilibrium** — the special case with correct beliefs.
- **2.6 Correlated equilibrium** — broader than Nash, incomparable to rationalizability.
- **4.3 Bayesian Nash equilibrium** — where beliefs over types become explicit.
- **Epistemic game theory** — formal treatment of common knowledge of rationality.

Structural Compression

Invariant: Rationalizable strategies are exactly those surviving iterated best-response elimination: a strategy survives at round $k+1$ iff it is a best response to some belief over opponents' round-k surviving strategies.

Mechanism: Iterative best-response filtering on the joint strategy space; for 2 players, equivalent to iterated strict dominance; for $n \geq 3$, equivalent under correlated beliefs.

Cross-System Echo: Rationalizability is the strategic analogue of "provable in at most k steps of a formal system": at round k, a strategy is rationalizable iff it is justifiable by a proof tree of depth k in the logic of rational best response. The limit is the set of strategies provable under unbounded common knowledge of rationality.

Exercises

1. In the Prisoner's Dilemma, show that both Defect profiles are rationalizable and Cooperate is not. Conclude that rationalizability gives the same answer as Nash in dominance-solvable games.
2. In Rock–Paper–Scissors, show that every pure strategy is rationalizable.
3. Prove Pearce's Lemma: a pure strategy s_i in a two-player finite game is never a best response iff it is strictly dominated by a mixed strategy.
4. Construct a 3-player game in which a strategy is rationalizable under correlated beliefs but not under independent beliefs.
5. Show that every strategy in the support of a Nash equilibrium is rationalizable.
6. Give an example of a rationalizable strategy that is not in the support of any Nash equilibrium.
7. **Argue, structurally,** why rationalizability is a weaker solution concept than Nash equilibrium in general but a stronger one than IESDS in the multi-player case (under independent beliefs). Identify which epistemic assumption — correctness of beliefs, independence of beliefs, or common knowledge of rationality — distinguishes each concept.

Game Theory · Module 2 — Nash Equilibrium · Node 2.5
Concepts: dominance, nash-equilibrium
Prerequisites: 1.3, 1.4, 2.1, 2.3
Enables: 2.6, 4.3

hbar.university — own.your.cognition