The Centipede Game and Common Knowledge

Game Theory · Module 3 — Extensive Form Games

Node 3.6

Position in Knowledge Graph

System: Game Theory **Structural Domain:** Extensive Form Games **Node:** 3.6

The Centipede game (Rosenthal, 1981) is a finite-horizon perfect-information game whose unique backward induction / SPE prediction is "Take immediately, at the very first stage." Empirically, people do not play this way — they pass for several rounds before taking, generating a collective surplus that SPE says should not arise. The Centipede is the paradigmatic puzzle of backward induction: it shows that the prediction of a well-defined solution concept can fail in experiments, and it motivates refinements involving bounded rationality, incomplete information about rationality, and common-knowledge failures. This lesson closes Module 3 by confronting the limits of equilibrium reasoning and motivating Modules 4 (incomplete information) and 5 (repeated games with reputation).

Formal Definition

Centipede game (Rosenthal, 1981). Two players alternate over n stages. At each stage $k \in \{1, \ldots, n\}$, the player whose turn it is decides to **Take** (T) or **Pass** (P).

- If Take at stage k, the game ends with payoffs that depend on k and who moves (specified below).
- If Pass, the pot doubles and the next stage begins.
- If all n stages Pass, the game ends with the maximum pot, split according to some fixed rule.

Concretely (standard version): at stage k, if the moving player Takes, they receive $2^{k-1} + c$ and the other receives $2^{k-1} - c$ for some $c > 0$. Passing doubles the pot. So the immediate Take is always marginally better than passing for the mover *at that specific stage*, because after passing, the opponent will Take next and leave the mover with less. But if both players passed throughout, the total surplus would be much larger.

SPE / backward induction prediction. Stage n: mover Takes (higher than passing leaves them). Stage $n-1$: knowing stage n mover will Take, mover at $n-1$ compares Take vs Pass (which leads to being Taken on by opponent); Take is better. Iterating: at every stage, Take is the unique SPE action. Unique SPE outcome: mover at stage 1 Takes immediately, receiving $1 + c$, other gets $1 - c$.

Empirical observation. McKelvey and Palfrey (1992) ran experiments: ~75% of pairs pass at the first stage, ~40% pass at the second, ~20% pass at the third, and about 1% make it to the last stage before one takes. Nobody plays "Take immediately."

Intuition

The Centipede game is strategically bizarre: the SPE prediction is Pareto-dominated by almost every other possible outcome. Both players would be better off if both Passed several times, but neither can credibly commit to Passing, because at any stage Taking is strictly individually rational (given the opponent will Take next).

The backward induction unravels from the last stage. You can't pass at the last stage because the game ends with the pot split badly for you. You can't pass at the second-to-last, because knowing the opponent will Take, you get less. Iterating, unraveling propagates all the way to stage 1, and you Take immediately.

Why experiments diverge. Several explanations compete:

1. **Bounded rationality.** Subjects cannot reason through all n steps of backward induction. Evidence: deeper centipedes have faster unraveling than shallow ones, inconsistent with full backward induction.
2. **Incomplete information about rationality.** If a player thinks there is even a small probability that the opponent is irrational (plays Pass even when Taking is better), then early Passing becomes rational: you Pass to "test" whether the opponent will Pass too, with the hope of a large future reward. Kreps–Milgrom–Roberts–Wilson (1982) formalized this in the context of the chain store paradox.
3. **Other-regarding preferences.** Players care about fairness or reciprocity, not just own payoff. Passing is cooperative, Taking is defection; the true payoff function includes a fairness term.
4. **Common knowledge failure.** Aumann (1995) showed that SPE requires common knowledge of rationality, and this is a strong assumption. Relaxing it allows more equilibria.

Any of these resolves the paradox. All are interesting. All motivate topics in later modules.

Structural Role

The Centipede game is the bridge between Module 3 (SPE) and Modules 4, 5, 10:

- **3.4, 3.5 $\Rightarrow$ 3.6.** SPE gives the unravel prediction.
- **3.6 $\Rightarrow$ 4.x.** Incomplete information about rationality is developed via Bayesian games.
- **3.6 $\Rightarrow$ 5.5 Chain store paradox.** The same unravel logic, with reputation as the mechanism to escape it.
- **3.6 $\Rightarrow$ 10.4 Commitment.** If players could commit to passing for k stages, both would do strictly better.
- **3.6 $\Rightarrow$ 9.x Learning.** Learning dynamics can converge to non-SPE equilibria, especially when common knowledge of rationality is relaxed.

Mathematical Development

Backward induction on the standard centipede

Consider $n = 6$ centipede with starting pot 2 and doubling. Payoffs at stage k (mover takes): mover gets $(3/4) \cdot 2^k$, other gets $(1/4) \cdot 2^k$. Passing doubles the pot.

Stage	Mover	Take payoff (mover, other)	Pass leads to
1	P1	(1.5, 0.5)	Stage 2 with pot 4
2	P2	(3, 1)	Stage 3 with pot 8
3	P1	(6, 2)	Stage 4 with pot 16
4	P2	(12, 4)	Stage 5 with pot 32
5	P1	(24, 8)	Stage 6 with pot 64
6	P2	(48, 16)	End with pot 128 split (96, 32)

Backward induction from stage 6: P2 compares Take (48) to Pass (32 from end) $\rightarrow$ Take.

Stage 5: P1 compares Take (24) to Pass leading to stage 6 where P2 Takes, leaving P1 with 16 $\to$ Take ($24 > 16$).

Stage 4: P2 compares Take (12) to Pass leading to stage 5 where P1 Takes (leaving P2 with 8) $\to$ Take ($12 > 8$).

Stage 3: P1 Take (6) vs Pass $\to$ P2 Takes at stage 4 leaving P1 with 4 $\to$ Take ($6 > 4$).

Stage 2: P2 Take (3) vs Pass $\to$ P1 Takes at stage 3 leaving P2 with 2 $\to$ Take ($3 > 2$).

Stage 1: P1 Take (1.5) vs Pass $\to$ P2 Takes at stage 2 leaving P1 with 1 $\to$ Take ($1.5 > 1$).

Unique SPE outcome: Take at stage 1. Payoff (1.5, 0.5). The reachable surplus (128) is never realized.

Pareto inefficiency

The SPE outcome dominates no other outcome — it has total surplus 2, while "both pass to the end" gives 128, with both players much better off than in SPE. The SPE outcome is Pareto-dominated by "pass all 6" and by "pass for several stages then take."

Kreps–Milgrom–Roberts–Wilson: tiny irrationality resolves the paradox

Add $\epsilon > 0$ probability that the opponent is an "irrational always-Pass type." Then Passing at stage 1 gives some chance of reaching stage 2 where the rational player might also Pass (to maintain the pretense of being irrational, for large future gains). The equilibrium of the perturbed game has substantial early passing, converging to the observed empirical behavior as $\epsilon \to 0$.

This shows the SPE of the full-rationality game is *not continuous in the common-knowledge assumption.* A tiny epistemic perturbation changes predictions drastically.

Worked Example

Empirical data from McKelvey–Palfrey (1992)

In a 4-stage centipede with increasing doublings: - ~80% of pairs Pass stage 1. - ~50% Pass stage 2. - ~25% Pass stage 3. - ~5% Pass stage 4 (reach the end).

SPE prediction: 0% Pass stage 1, 100% Take. The gap is dramatic.

Interpretation. Players cooperate for several rounds, generating surplus, until one defects. The modal outcome is not "Take immediately" but "Take around stage 2 or 3." Replicated across many experimental studies and subject populations.

Quantal response equilibrium explanation

A quantal response equilibrium (McKelvey–Palfrey, 1995) is a "logit"-smoothed Nash equilibrium where players don't best-respond deterministically but with probabilities proportional to $e^{\lambda \cdot \mathrm{EV}}$ for some rationality parameter λ. For small λ (bounded rationality), the equilibrium of a centipede game involves substantial early passing, matching experimental data.

This is a formal model of deviations from SPE. As $\lambda \to \infty$, QRE $\to$ SPE. For finite λ, QRE predicts the Pass-heavy early-stage behavior observed experimentally.

Poker Anchor

Concept mapping

Formal object	Poker instantiation
Centipede unraveling	The theoretical argument that in a pot with multiple streets of betting, one should fold earlier than feels right
SPE prediction vs empirical play	GTO "folds too much" for humans; humans "pass" (call down light) beyond what solver recommends
Irrationality epsilon	Uncertainty about opponent's type (fish, thinking player, GTO bot)
Pareto-dominated equilibrium	The SPE prediction gives lower EV than mutually passive play

Concrete situation

Multi-street pot with polarized opponent. You hold a mid-strength hand (e.g., top pair with weak kicker) and are facing bets on every street. The solver says to fold on the turn (because opponent's turn barrel is too polarized — mostly value, not enough bluffs). You call anyway, figuring you might catch a bluff on the river.

The centipede analogy. Each street is a stage. "Call" is roughly "Pass" (continue to next stage); "Fold" is roughly "Take" (end the game with your current EV). GTO says fold earlier than feels right, because the opponent's betting range gets more polarized on each street, and your bluff-catcher is "unraveled" from the river back.

Empirical bias. Most players hate folding, especially on turn and river, because the immediate impulse is "I already have money in the pot." The SPE prediction (fold the turn) is correct under equilibrium assumptions, but if the opponent is irrational in a specific way (over-bluffs), then calling is better. The Kreps–Milgrom–Roberts–Wilson resolution applies: if you believe there's a non-trivial probability the opponent is over-bluffing (many live recreational players do), then passing (calling) has positive EV despite the SPE prediction.

Trust in backward induction vs empirical adaptation. A strong GTO-trained player folds according to the equilibrium prediction even when it feels wrong; a weaker player "hero-calls" based on the suspicion that the opponent is irrational. Both can be correct — the question is whether the epsilon is actually there.

What this understanding buys you

The bluff-catcher dilemma is a centipede in miniature. GTO tells you to fold more than emotionally feels right. Empirically, most amateur opponents bluff less than GTO-optimal frequency, so the right answer is to fold *even more* than GTO says. But the opposite mistake — hero-calling against a polarized range — is also common. The centipede framing says: trust the backward induction, and fold when you should fold.

Reputation matters across sessions. In a single hand, you play the equilibrium of the hand. Across hands, your history affects opponent beliefs. Playing a non-SPE "always call down" strategy builds a reputation that opponents exploit by bluffing less and value-betting thinner — exactly the KMRW dynamic. Module 5.5 (chain store paradox) makes this precise.

Common knowledge failures are normal. You do not actually know whether your opponent knows you know they know about GTO. This is a chain of beliefs that, at some level, fails. The centipede game is the formal demonstration that when this chain is short, the predictions of full-rationality game theory diverge from reality.

Cross-Links

- **3.4 Backward induction** — the computation that produces the Centipede SPE.

- **3.5 Subgame perfection** — the refinement that uniquely pins down Take-immediately.
- **4.3, 4.4 Bayesian Nash, PBE** — how incomplete information rescues the empirical predictions.
- **5.5 Chain store paradox** — reputation in a similarly unraveling game.
- **5.6 Cooperation in repeated PD** — empirical cooperation vs SPE prediction.
- **10.4 Commitment** — Pareto improvements from commitment devices.
- **QRE, bounded rationality** — formal models of the gap.

Structural Compression

Invariant: In a finite-horizon perfect-information game with unilaterally dominant "defection" actions at each stage, backward induction unravels all cooperation, producing a unique SPE that is Pareto-dominated by almost every other outcome. Empirically, this prediction fails.

Mechanism: Backward recursion from the terminal stage through each decision node, with the common-knowledge-of-rationality assumption propagating the unraveling all the way to the first stage.

Cross-System Echo: The centipede game is the game-theoretic instance of a broader pattern: "logically forced defection" appears in the chain store paradox, finitely repeated prisoner's dilemma, public goods provision, and pollution games. In each case, the equilibrium unraveling argument gives a Pareto-dominated prediction, and experimental subjects routinely fail to unravel. The discrepancy is a pointer to the limits of common-knowledge reasoning.

Exercises

1. Verify the backward induction of a 4-stage centipede with payoffs $(1, 0.5), (0.5, 2), (4, 1), (2, 8)$, and terminal split $(16, 16)$.
2. Show that every stage of a generic centipede game has Take as the unique SPE action.
3. Compute the payoff gap (SPE vs "Pass all") for a centipede of length n with a doubling rule.
4. In the KMRW setup with ϵ irrationality probability, derive the first stage at which Taking becomes optimal.
5. Compute the quantal response equilibrium of a 4-stage centipede for $\lambda = 1$. Does it predict substantial early passing?
6. Give an example of a perfect-information game where SPE coincides with an intuitive cooperative outcome rather than defecting. What structural feature distinguishes it from the centipede?
7. **Argue, structurally,** why the Centipede game is a proof that backward induction requires common knowledge of rationality, not just individual rationality. Identify the exact point in the unraveling argument that fails if one player only knows they are rational but is not sure the opponent knows they know.

Game Theory · Module 3 — Extensive Form Games · Node 3.6
Concepts: nash-equilibrium, repeated-interaction
Prerequisites: 3.4, 3.5
Enables: 5.5, 10.4

hbar.university — own.your.cognition