Testing Backward Induction

Summer 26

Niels Boissonnet, Bielefeld

Discussion

A dichotomy

- Social preferences. Deviations from Nash and SPNE are pervasive (e.g. in ultimatum games).

  Two broad categories of explanation:

  - Replace "payoffs" with "utilities": inequity aversion, ERC, reciprocity.
  - Replace "equilibrium" with "bounded strategic thinking": level-k, cognitive hierarchy, QRE.

- Today's lecture. *Backward induction* itself — the cognitive engine of subgame perfection. We separate two questions usually conflated:

  1. Do people have the preferences classical theory assumes?
  2. Do people *reason* backwards from the end of the game?

  Even granting the first question to behavioural theory, the second remains open.

Learning goals for this lecture

By the end of this lecture you should be able to:

1. State subgame consistency and truncation consistency and explain why they follow from any (rational) preference-based account that uses backward induction.

2. Describe the Binmore (3) evidence that each of these conditions is violated — and the "rescue" stories the paper rules out.

3. Describe the Johnson et al. (15) evidence — from process tracing and from human-vs.-robot matchings — that points to *limited look-ahead* rather than fairness as the binding constraint.

4. Place these results in the broader literature on bounded strategic reasoning: level-k models, cognitive hierarchy, and the more recent process-tracing and centipede-game literature.

Backward induction

In a finite extensive-form game of perfect information:

1. At each terminal decision node, the player chooses an action maximising their (own) utility.
2. Replace each such decision node with the resulting equilibrium payoff vector.
3. Repeat, working backwards to the root.

This procedure delivers the (generically) unique subgame-perfect Nash equilibrium (SPNE) of the game (23; 17).

Why it matters. BI is the textbook justification of SPNE.

What it presupposes.

- rationality at every node, including off-path ones,
- common knowledge of rationality (2),
- consistent preferences over outcomes, and
- the cognitive capacity to look ahead and back-substitute.

The empirical puzzle

Decades of ultimatum find:

- Offers around 40–50% of the pie, not the SPNE offer near zero.
- Substantial rejection of unequal offers, even at material cost.
- Similar deviations in trust games, gift-exchange, dictator games.

The preference rescue. Subjects do backward-induce, but on *social* preferences:

- Fehr and Schmidt (12): Inequity aversion
- Bolton and Ockenfels (4): utility depends on own payoff and own share.
- Rabin (22): utility depends on intentions.

This does not confirm backward induction. If we only observe choices, preferences and reasoning are jointly identified.
$\Longrightarrow$ We can fit nearly any behavior with the right preferences — so "do people backward-induce?" is not really being tested.

Today's two strategies for testing BI directly

1. Binmore (3): Test behavioural implications of BI for *any* preferences.
   $\Longrightarrow$ isolate two such implications — subgame and truncation consistency and show both are experimentally violated.

2. Johnson et al. (15): Open the black box of the decision process.
   $\Longrightarrow$ use mouse-tracking to record *what information subjects look at*
   $\Longrightarrow$ replace human opponents with robots that play SPNE.

Both papers were published in the same 2002 issue of *JET* and have become standard references for behavioural game theory's claim that BI fails as a positive theory of choice in extensive-form games.

Roadmap

1. Overview & motivation (you are here)
2. Binmore (3): design & subgame consistency
3. Binmore (3): truncation consistency & conclusions
4. Johnson et al. (15), study 1: MouseLab and cognitive process
5. Johnson et al. (15), study 2: robots & training
6. Synthesis & recent literature

Next (Section 2): The first paper. We will see how a clever pair of game-comparisons can identify BI failures without committing to any particular preference specification.

Subgame Consistency

Testing Backward Induction

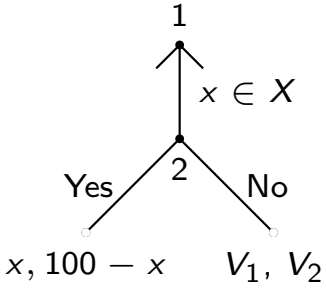
Method: three conditions for "backward induction"

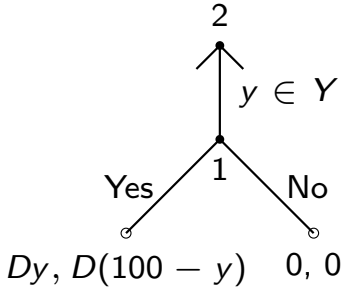
Any account in which players backward induce must satisfy:

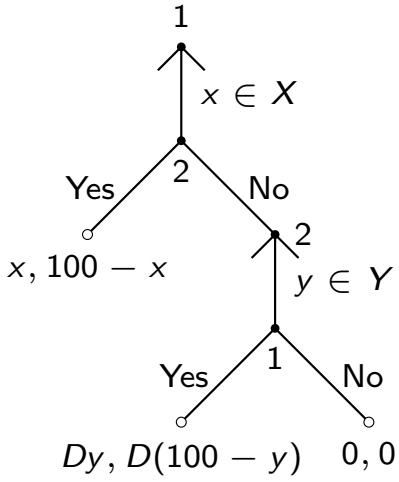
- Rationality: Given a choice between two (vectors of) payoffs, a player picks the most preferred.
- Subgame Consistency: Play in a subgame is independent of the subgame's position in a larger game.
- Truncation Consistency: Replacing a subgame with its equilibrium payoffs does not affect play elsewhere in the game.

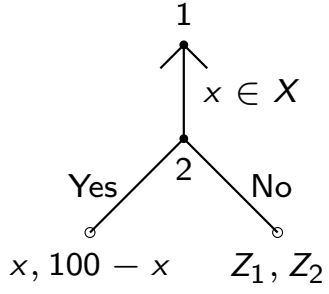
Assuming Rationality, and test Subgame Consistency and Truncation Consistency.
$\Longrightarrow$ Violations of either are violations of BI regardless of what preferences look like.

The four games









I: measures (V_1, V_2)

II: continuation game

III: two-stage bargaining

IV: truncated III

Roles of the games.

- Game I elicits "rejection valuations" (V_1, V_2).
- Game II is the standalone continuation. Discount factor $D \in \{.2, .3, \ldots, .9\}$.
- Game III is the full two-stage game.
- Game IV is Game III with the continuation replaced by fixed payoffs (Z_1, Z_2).

Subgame consistency: the comparison

"Play in a subgame is independent of the subgame's position in a larger game" (compare Game II with the second stage of Game III).

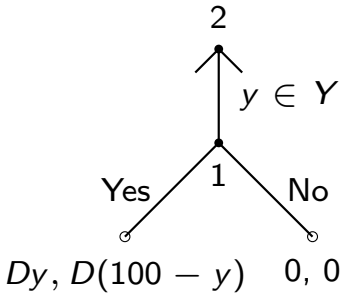


Figure: Game II

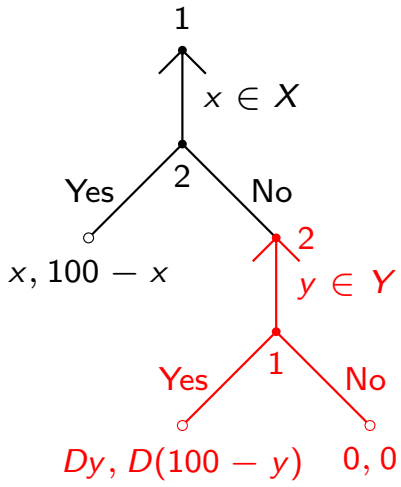


Figure: Game III

Under any payoff-interdependent preferences, BI predicts that the red subtree is played the same way as the standalone Game II.
Reciprocity (22), however, makes play depend on what happened earlier — violating subgame consistency by construction.

Is subgame consistency violated?

Isn't the mere fact that stage 2 is reached evidence that BI fails?

- No. Players' (interdependent) preferences are not commonly known, so positive-probability rejection can be SPNE behaviour.

Does a difference between player-2 offers in stage 2 of Game III and in Game II indicate a violation?

- Not necessarily. Player 2 may have updated about player 1's preferences from the demand 2 just rejected.

So the empirical question is sharper: can we explain any observed discrepancy by belief updating that is itself consistent with BI?

- If yes (e.g. 2's stage-2 offer co-varies cleanly with 1's stage-1 demand), the violation evaporates.
- If no, the violation is real.

Headline evidence: offers differ substantially

Rounds	(V_1, V_2)	Observations	Mean offer	Median	5th %tile	95th %tile
1-10	All	1200	33.4	33	15	50
11-20	All	1200	32.2	31	20	48.5

Figure: Game II

Specification	Rounds	Observations	Mean offer	Median	5th %tile	95th %tile
All	1-10	276	43.3	44	23	60
All	11-20	225	43.8	45	25	60
All	1-20	501	43.5	44	25	60

Figure: Second stage of Game III

...and rejection rates are higher too

		Offers		
Game	Observations	All	30–40	35–45
II	2400	20%		
III	501	29%		
II	1064		15%	
III	131		47%	
II	733			9%
III	190			36%

Rejection rates for the continuation game (Game II) and the second stage of the two-stage game (Game III). 2400 plays of each; 501 reached stage 2.

- Higher rejection rates in stage 2 of Game III show that the higher offers *were necessary* — but the violation of subgame consistency is precisely this packet of (offer, rejection) behaviour being different.
- Even for offers in the same numerical range (30–40 or 35–45), rejection in stage 2 of Game III is 2–4$\times$ higher.

Effect of the discrepancy on expected payoffs

Game	D	II-median, y_m^{II}	Payoff	III-median, y_m^{III}	Payoff
II	.2	31	13	45	11
II	.3	28	19	47	16
II	.4	35	23	48	21
II	.5	33.5	33	44	28
II	.6	35	34	37.5	33
II	.7	31	41	45	39
II	.8	35	52	40.5	48
II	.9	25	68	45	50
III	.2	31	0	45	7
III	.3	28	0	47	13
III	.4	35	0	48	9
III	.5	33.5	0	44	18
III	.6	35	32	37.5	38
III	.7	31	0	45	39
III	.8	35	15	40.5	27
III	.9	25	0	45	41

Game-II median offers would earn very little if made in stage 2 of Game III, and vice versa. Player 2 cannot "just play the Game-II strategy" once stage 2 of Game III is reached.

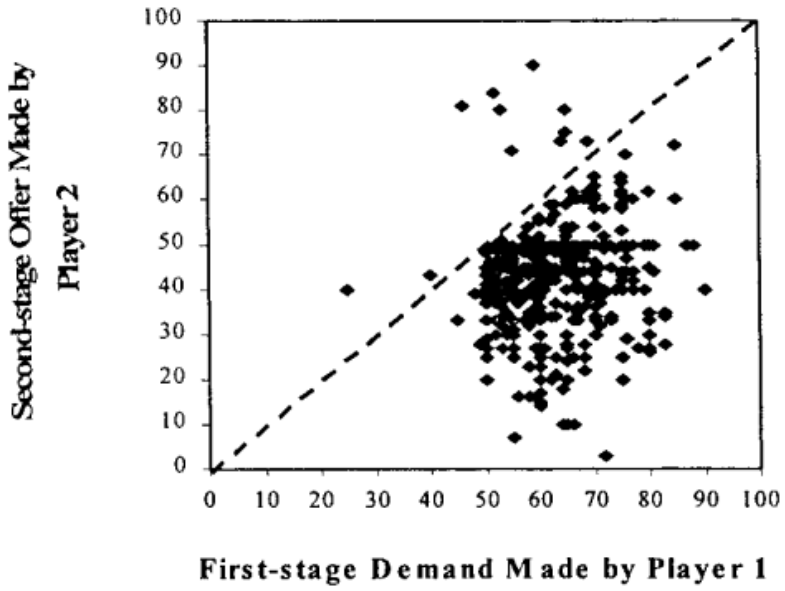
Two candidate explanations

1. Belief update about the opponent's preferences. Proposers in stage 2 of Game III may simply be reacting to information about their opponent's preferences contained in 1's first-stage demand.

2. Selection. Stage 2 of Game III is reached only if player 2 rejects 1's initial demand. Maybe the player-2s who reject are an atypical sample (e.g. more fairness-minded).

If either holds, the apparent inconsistency disappears. BMPSS test both.

Belief update (1/2): is the stage-2 offer informative?



- If player 2 inferred something useful about player 1's preferences from the first-stage demand, 2's stage-2 offer should co-vary systematically with that demand.
- The scatter shows essentially no relationship between the rejected demand and the subsequent offer.

Belief update (2/2): is there anything to learn?

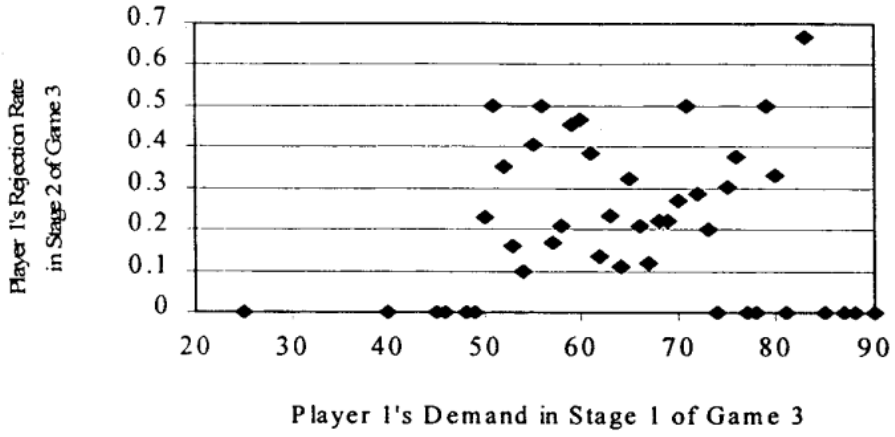


FIG. 14. Player 1's rejection rate in the second stage of game III, as a function of the first-eriod demand that player 1 had rejected in order to reach the second stage.

- Player 1's stage-2 rejection rate is essentially *flat* in the first-stage demand 1 made.
- There is little for player 2 to learn from observing 1's first-period demand — and indeed 2 does not appear to learn from it.
- Decomposing by discount factor and controlling for 2's stage-2 offer (small samples) yields similar results.

Selection: is the stage-2 sample atypical?

Mean offer in Game II — restricted to player-2s who rejected at least one demand in Game III

Rounds	(V_1, V_2)	Observations	Mean offer	Median	5th %tile	95th %tile
1-10	All	777	33.5	33	16	50
11-20	All	780	32.4	31	20	49

Mean offer in Game II — full sample

Rounds	(V_1, V_2)	Observations	Mean offer	Median	5th %tile	95th %tile
1-10	All	1200	33.4	33	15	50
11-20	All	1200	32.2	31	20	48.5

The two distributions are essentially identical (33.5 vs. 33.4 mean in rounds 1–10; 32.4 vs. 32.2 in rounds 11–20). Selection is not the story.

Section 2 take-aways

- Backward induction $\neq$ "play the SPNE"
  Binmore (3): it requires rationality, subgame consistency, and truncation consistency.

- Subgame consistency fails.
  Stage-2 offers are higher than in Game II, with more rejections.

- Why? Not explained by belief updating or player-2 selection.

- Leading candidate: reciprocity.
  Rejecting puts player 2 in a different psychological state — and reveals player 1's "unfairness" (22).

- Next (Section 3): Testing *truncation consistency*.

Truncation Consistency

Testing Backward Induction

Truncation consistency: the comparison

"*Replacing a subgame with its equilibrium payoffs does not affect play elsewhere in the game*" (Game IV vs first stage of Game III).

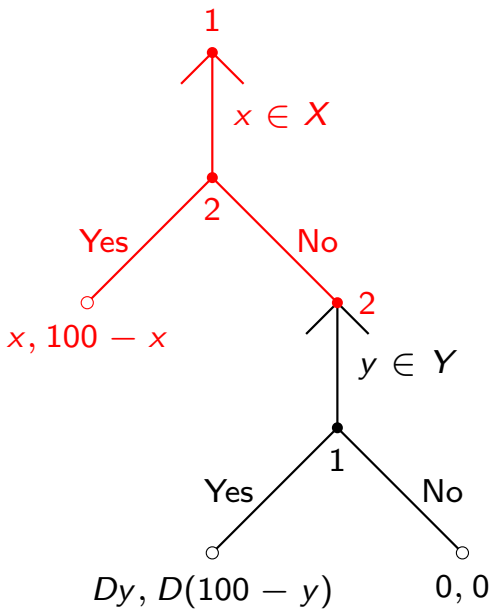


Figure: Game III

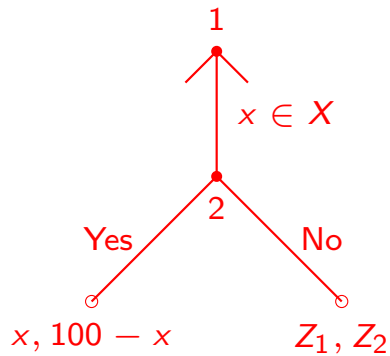


Figure: Game IV

Under any payoff-interdependent preferences, BI predicts that stage-1 behaviour in Game III equals behaviour in Game IV when (Z_1, Z_2) matches the expected stage-2 payoffs of Game III.

How (Z_1, Z_2) was determined

BMPSS set $(Z_1(i), Z_2(i))$ *individually*: for each participant i.

- $Z_1(i)$ = mean of P1's payoffs in second-stage games in which i participated.
- $Z_2(i)$ = mean of P2 payoffs's in second-stage games in which i participated.

Why this works. If players have payoff-interdependent preferences and use BI, then

- in Game III they should fold the second stage into its certainty-equivalent
- Under the empirical (Z_1, Z_2), that certainty-equivalent is exactly whenever the player is risk neutral.

Note. This is a fairness-robust comparison: any utility function over outcomes would yield the same prediction in Game III stage 1 and Game IV.

What truncation consistency predicts

- Player-1 demands in Game III stage 1 should depend strongly on D, because D determines the value of continuing.
- Demands should depend on (Z_1, Z_2) in Game III and on (Z_1, Z_2) in Game IV in the same way.

Quantitatively (theoretical). For a wide range of preferences, as D rises from .2 to .9, predicted demands fall from ≈ 80 to ≈ 10 (player 1 becomes weak as the pie shrinks slowly).

Result: Game-III demands are barely sensitive to D

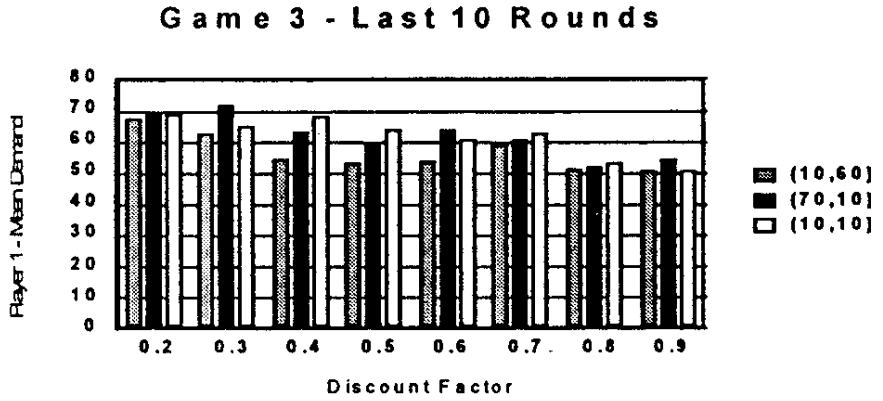


FIG. 16. Mean and median demands made by player 1 in the first stage of the two-stage game, measured as a percentage of the surplus.

- Player-1 mean demands in Game III, by D and by previous Game-I rejection payoff $(V_1, V_2) \in \{(10, 10), (70, 10), (10, 60)\}$.
- Demands move only mildly from ≈ 67 at $D = .2$ to ≈ 52 at $D = .9$ — far from the predicted swing from 80 to 10.
- Past-experience effects are small; the predicted strong response to the value of continuation is largely absent.

Game III is less sensitive than Game IV

Independent variable	Estimated coefficient	Standard error	p-value
Intercept	.63	.027	.000
Player-1 payoff	.030	.0011	.000
Player-2 payoff	-.027	.00058	.000
Player-1 payoff, Game III	-.035	.0017	.000
Player-2 payoff, Game III	.025	.0012	.000
(70,10)	.024	.031	.44
(10,60)	-.21	.032	.000

FIG. 17. Random effects estimates of player 1's first-stage demand in Games III and IV. The dependent variable is $\log x/(100-x)$, where x is player 1's demand in the first stage of game III or game IV. The player-1 and player-2 payoffs are (Z_1, Z_2) for game-IV observations and (Z_1^{III}, Z_2^{III}) for game-III observations. "Player-1 payoff, Game III" and "Player-2 payoff, Game III" are dummy variables, indicating how the coefficients on player-1 and player-2 payoffs for game III differ from the game-IV base case. The game-I rejection payoff (10, 10) is the intercept base case, with dummy variables for the (70, 10) and (10, 60) cases.

- Player-2 payoff. Larger Z_2 in Game IV $\Rightarrow$ more moderate demands by 1.
- Player-2 payoff, Game III dummy. The dependence on Z_2 is much weaker in Game III than in Game IV.
- Player-1 payoff. Larger Z_1 in Game IV $\Rightarrow$ more aggressive demands by 1.
- Player-1 payoff, Game III dummy (negative). Again, the sensitivity is attenuated in Game III.

Punchline. Players respond to *fixed* continuation payoffs (Game IV) more strongly than to *stochastic* continuation games with the same expected payoffs (Game III).

Why might truncation consistency fail?

Two leading explanations from the literature:

- Strategic / ambiguity aversion. The continuation in Game III involves uncertainty about what the opponent will do; a fixed (Z_1, Z_2) does not. If players are averse to strategic uncertainty (14), the two situations are not equivalent for them.
- Imperfect understanding / limited look-ahead. Subjects may simply not work out what the continuation is worth, treating it as a vague "something happens later" rather than as the precise lottery the experimenter has constructed. This anticipates the Johnson et al. (15) story.

Both explanations sit alongside reciprocity (from Section 2). They are not mutually exclusive.

Section 3 take-aways

- Truncation consistency is violated. Player-1 demands respond robustly to player-2 rejection payoffs in Game IV but much less in Game III.
- This is a violation of BI that does not require any commitment to a particular preference specification — any rational player would treat the two games as equivalent.
- Two viable explanations: (i) *strategic / ambiguity aversion* (14), (ii) *limited look-ahead*.
- Combined message of Binmore (3). The classical BI engine fails on two distinct dimensions; preference-based fixes alone are insufficient.
- Next (Section 4): Open the black box and watch the process.

Cognitive process via Mouselab

Testing Backward Induction

Where we are

- Sections 2–3 (recap): Binmore (3) infer BI failures from *outcomes*. Two candidate explanations remain — reciprocity and limited look-ahead — and outcomes alone cannot distinguish them.
- This section: Johnson et al. (15) *measure the cognitive process directly*. If players really backward-induce, the order and duration of information acquisition should look very different from forward-looking, fairness-driven choice.
- Up next (Section 5): Robots, to remove fairness motives entirely.

Learning goals. Understand MouseLab; recall the level-k classification; interpret information-search data as evidence on look-ahead.

From outcomes to process: a method swap

Outcome-based experiments treat the subject as a black box. Process-tracing experiments (8) ask: *what is the decision-maker actually doing*?

Tools available to behavioural game theorists:

- MouseLab. Information boxes subjects must actively open with a mouse. Used in Johnson et al. (15) and many follow-ups (Costa-Gomes et al. (8); Costa-Gomes and Crawford (7); Brocas et al. (5)).
- Eye-tracking. More natural, finer-grained Wang et al. (24); Polonio and Coricelli (21).
- Process measures (response times, retrospective reports, neural data).

Key prediction of BI. A backward inductor should look most at the *last* stage and make many backward transitions, computing the continuation value before deciding what to demand in stage 1.

The three-stage bargaining game

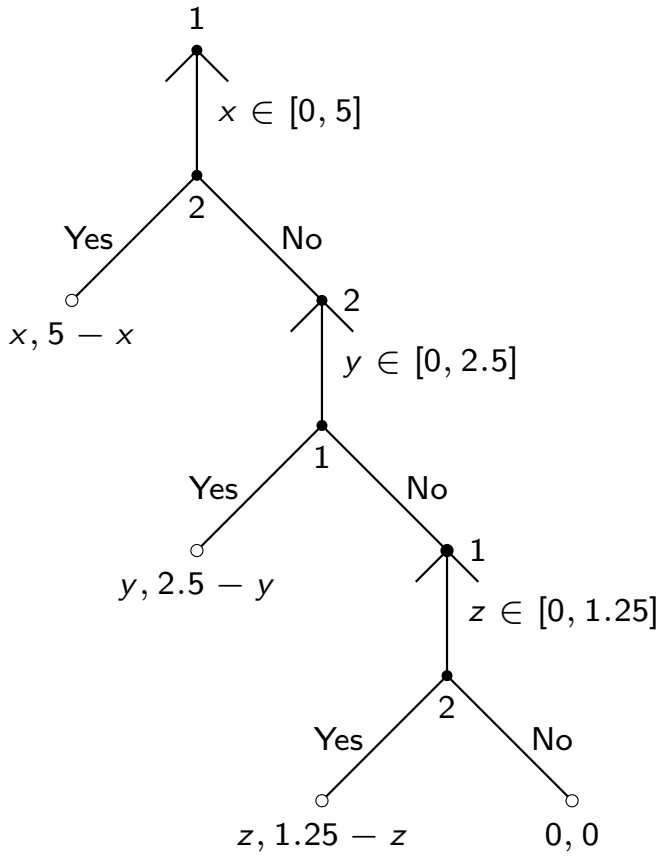


Figure: Three-stage bargaining

- Three alternating offers; pie shrinks by $\delta = .5$ each round.
- SPNE offer in round 1: \$1.25.
- Information is partitioned into three labelled boxes — one per round's pie size — which subjects must *open* to view.
- Subjects' opponents are computers preprogrammed to play SPNE-consistent rejections.

The Mouselab display

- Subjects move the mouse cursor *into* a box to see its contents.
- Every 16 ms the system records which box is open.
- Output. Per decision: the sequence of boxes opened, the duration each was open, transitions between boxes.
- Subjects make an offer by moving the cursor to a dollar scale at the bottom.

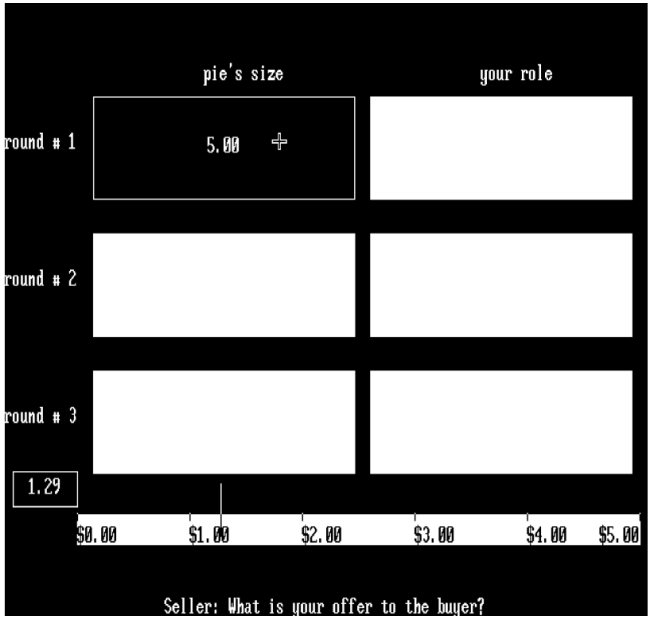


Figure: MOUSELAB computer display

Why this is a real test. A genuine backward inductor must look at round 3, then round 2, then round 1. A subject who only looks at round 1 either is not doing BI, or has memorised the answer.

What look-ahead types should we expect to see?

Limited look-ahead. A type-k player thinks k steps ahead and "starts" the backward induction there.

- Level 0. Treats round 1 as an ultimatum game. Predicted offer: $0.4 \times \$5 = \$2$.
- Level 1. Looks one round ahead; treats round 2 as the final ultimatum. Predicted offer: \$1.50 to player 2 in round 1.
- Equilibrium (Level ≥ 2). Reaches the SPNE offer of \$1.25.

Same logic as level-k in normal-form games (20; 9), here applied to extensive-form reasoning. Each level predicts a *specific search pattern*, not just an offer.

Results: offers and rejection rates

- Equilibrium offer: \$1.25
- Median offer in experiment: \$2.11 — close to the level-0 prediction.
- Offers below \$1.80 rejected $\geq 50\%$ of the time.
- Players look most often at the first-round pie size; least at round 3.

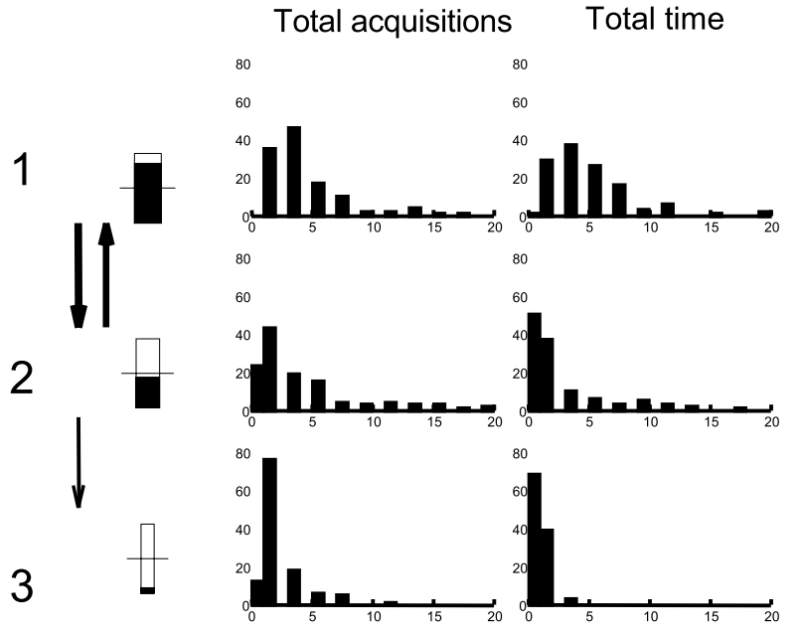
TABLE I

Round	Number of acquisitions	Total time examining payoff	Transitions		
			1	2	3
Round 1 (\$5.00)	4.38	12.91	—	2.55	0.65
Round 2 (\$2.50)	3.80	6.67	2.10	—	1.24
Round 3 (\$1.25)	2.12	1.24	0.50	0.88	—

- Forward transitions ($1 \to 2$, $2 \to 3$) outnumber backward ones.

$\Rightarrow$ No backward induction; no equilibrium behaviour.

Distribution of information-search effort



Histograms of number of acquisitions (left) and total inspection time (right) by round.

- Round 1 attracts the most acquisitions and time — long right tail of intensive lookers.
- Rounds 2 and 3 looked at little; most subjects barely open the round-3 box.
- The few who do look extensively at later rounds match the "equilibrium" look-ahead type and indeed make lower offers.

Linking process to outcomes

Johnson et al. classify subjects by their search pattern and correlate with offer behaviour:

- Level-0 lookers (mostly round-1 box) offer *near $2*.
- Level-1 lookers (rounds 1 and 2, forward transitions) offer $1.50–$2.
- Equilibrium lookers (extensive round-3 inspection, backward transitions) offer *near $1.25*.

Diagnostic value. Process data identifies the cognitive type without committing to a preference specification. A fairness-driven backward inductor and a level-1 player both offer more than SPNE — but they look very different on the inside.

Similar process-cognition mappings: Costa-Gomes et al. (8); Costa-Gomes and Crawford (7) for guessing games; Brocas et al. (5) for private-information games.

Section 4 take-aways

- Process data confirms BI failure much more sharply than outcome comparisons: most subjects do not look at the information they would need to backward-induce.
- Limited look-ahead is a positive theory: a substantial share looks one or two rounds ahead at most, makes forward transitions, and offers what level-0 or level-1 would predict.
- A critic can still object: maybe subjects *know* the SPNE but offer more because they expect humans to reject low offers. The next study tackles this.
- Next (Section 5): Replace the opponent with a robot that plays SPNE. If fairness was binding, offers should drop to \$1.25.

Study 2 — Robots, training, and what they reveal

Testing Backward Induction

Where we are

- Section 4 (recap): Subjects' information search is forward, anchored on round 1, and rarely reaches round 3.
- Open objection. Maybe subjects *know* the SPNE but offer more because they expect a human to reject "unfair" offers. Search patterns reflect strategic adaptation to human fairness, not failure of cognition.
- This section. Study 2 cleanly identifies fairness vs. cognition by replacing the human opponent with a robot that plays SPNE — and then by *teaching* subjects backward induction.
- Up next (Section 6): Synthesis and recent literature.

Why robots?

- Removes the fairness motive on the opponent's side. Subjects are told the computer maximises its own dollar payoff and assumes the human does too.
- Pins down beliefs. Subjects know the opponent's strategy precisely — no inference about preferences is needed.
- Leaves only the human's own motives and cognition.

Two clean predictions.

- Fairness-only. Behaviour should converge near SPNE.
- Limited-lookahead. Subjects who don't backward-induce will not *compute* SPNE and will keep offering high.

Design of Study 2

Three phases:

- Phase A: No instruction. Bargaining game from Study 1 (SPNE = \$1.25). Robot rejects offers below SPNE; subjects learn from feedback.
- Phase B: Instruction, same game. Subjects taught the backward-induction algorithm (without being told the answer).
- Phase C: Instruction, harder game. New parameters change the SPNE to \$1.51. If subjects internalised BI, they should transfer the algorithm.

Identifying assumption. Robots remove the strategic source of generosity; training removes the cognitive source. Comparing A, B, C identifies which was binding.

Results

TABLE II

Results in Training Study 2

	No instruction, equilibrium = $1.26				Post instruction, equilibrium = $1.26		Post instruction equilibrium = $1.51	
Round	1–2	3–4	5–6	7–8	9–10	11–12	13–14	15–16
Average offer	1.83	2.02	1.67	1.82	1.21	1.23	1.51	1.54
Number of rejections (out of 30)	25	11	12	13	1	4	6	0

- Phase A (no instruction): Offers around $1.7–$2.0 — only modestly lower than against humans, far above SPNE.
- Phase B (post-instruction): Offers drop sharply to about $1.21–$1.23.
- Phase C (new game, SPNE = $1.51): Offers move to $1.51–$1.54 with no rejections — subjects applied the algorithm to the new game.

What the trajectory tells us

- Fairness alone cannot account for high pre-training offers. Against a robot, offers stay near \$1.7–\$2.0 — too high if the only deviation were modest own-guilt.
- Cognition is the binding constraint. Brief training in BI collapses offers onto SPNE, even though the experimenters do not tell subjects what SPNE is.
- Subjects internalise a procedure, not an answer. Phase C uses a new game; trained subjects play it correctly.
- Caveat. Training works because the algorithm is short and transparent. In richer environments (deeper centipede, repeated games) training may not be enough (13; 18).

Section 5 take-aways

- Fairness vs. cognition is separately identified. Robots don't get subjects to play SPNE; training does. The gap was driven primarily by *cognitive limitation*.
- Limited look-ahead is a positive theory. Once taught, subjects apply the algorithm correctly to transformed games.
- Combined message of Johnson et al. (15). The BI failure of Binmore (3) can be traced to a specific cognitive shortcut. Behaviour is well-described by a small level-k classification, augmented with mild own-fairness preferences.
- Next (Section 6): What does the broader literature look like, and what has happened since 2002?

Synthesis & Recent Literature: What have we learned, and where does the field stand?

Testing Backward Induction

Where we are

- Sections 2–3 (Binmore): *subgame* and *truncation* consistency are both violated.
- Sections 4–5 (Johnson et al.): process and intervention evidence point to limited look-ahead as the binding constraint.
- This section. Synthesis. Where does the literature stand 20+ years on?

Learning goals. Articulate the joint message; place results in level-k / cognitive-hierarchy and centipede literatures; understand when BI is and is not a reasonable approximation.

The joint message of the two papers

- Behavioural game theory inherited from Selten: *behaviour* = *preferences* + *backward induction*. Decades of work refined the "preferences" side.
- Binmore (3); Johnson et al. (15) push back on the other side. Even with arbitrary preferences, BI's behavioural fingerprints — subgame consistency, truncation consistency, backwards-then-forwards information search — are not in the data.
- Implication. Treat BI as a special case of a broader cognitive process, not the default. Two complementary modelling traditions follow.

Tradition 1: Cognitive hierarchies and level-k

- Level-k (**?** 20; 9): players best-respond to a lower-level opponent.
- Cognitive hierarchy (6): each level best-responds to a mixture of all lower levels.
- In extensive-form games this gives *limited look-ahead* — exactly the Johnson et al. interpretation.
- The distribution of k is fairly stable across games: few level-0s, many level-1/2s, few equilibrium types (9).

Recent developments. Structural estimation of look-ahead in dynamic games (5); level-k in centipede (**?** 13).

Tradition 2: Reciprocity and intentions

- For violations of *subgame* consistency, level-k alone is insufficient — stage 2 of Game III is played differently from a structurally identical Game II.
- Models of sequential reciprocity (22; Dufwenberg and Kirchsteiger; 11) extend utilities to depend on the *intentions* signalled by past play.
- Direct empirical support: **?**); axiomatisation in **?**).

Reciprocity and limited look-ahead are not competitors — they address different failures. A complete model of two-stage bargaining plausibly needs both.

Generality: when does BI work?

Twenty years of follow-up suggests BI is reliable when:

- The tree is short and the algorithm is short.
- Subjects are experienced or trained — even briefly (15**?**).
- Stakes are high enough to justify cognitive effort (1).
- There is no salient fairness motive.

And less reliable when:

- The tree is deep (centipede with 6+ nodes — 19**?**).
- Players have other-regarding motives or strategic uncertainty.
- Information costs of look-ahead are high.

Natural experiment. **?**) find that chess players — trained at backward induction — play centipede much closer to SPNE. The skill is real, learnable, and rare.

Process tracing has matured

- MouseLab is now standard (8; 7; 5).
- Eye-tracking provides more natural data (16; 24; 21).
- Studies routinely find *how people look* predicts *how they play* better than economic theory predicts from incentives alone.
- The methodology now travels: guessing games, signalling games, dynamic public-goods games.

Take-home for the course

- Backward induction is a strong cognitive assumption, not a benign mathematical convenience. Treat it as a hypothesis to be tested.
- Preference-based and cognition-based extensions are complementary. A workable applied model often combines a fairness preference with limited look-ahead.
- Method matters. What we conclude about a deviation depends on what we measure: outcomes, processes, or both. The Johnson–Camerer agenda is now standard.
- Applied implication. In mechanism and market design, "the SPNE will be played" is a strong claim. Design with bounded-reasoning agents in mind, or train them.

References I

[1] Andersen, S., Ertaç, S., Gneezy, U., Hoffman, M., and List, J. A. (2011). Stakes matter in ultimatum games. *American Economic Review*, 101(7):3427–39.

[2] Aumann, R. J. (1995). Backward induction and common knowledge of rationality. *Games and Economic Behavior*, 8(1):6–19.

[3] Binmore, K. (2002). Backward induction and rationality. Unpublished manuscript.

[4] Bolton, G. E. and Ockenfels, A. (2000). Erc: A theory of equity, reciprocity, and competition. *The American Economic Review*, 90(1):166–166.

[5] Brocas, I., Carrillo, J. D., Wang, S. W., and Camerer, C. F. (2014). Imperfect choice or imperfect attention? understanding strategic thinking in private information games. *Review of Economic Studies*, 81(3):944–970.

[6] Camerer, C., Ho, T.-H., and Chong, J. K. (2004). A cognitive hierarchy model of games. *Quarterly Journal of Economics*, 119(3):861–898.

[7] Costa-Gomes, M. A. and Crawford, V. P. (2006). Cognition and behavior in two-person guessing games: An experimental study. *American Economic Review*, 96(5):1737–1768.

[8] Costa-Gomes, M. A., Crawford, V. P., and Broseta, B. (2001). Cognition and behavior in normal-form games: An experimental study. *Econometrica*, 69(5):1193–1235.

References II

[9] Crawford, V. P., Costa-Gomes, M. A., and Iriberri, N. (2013). Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications. *Journal of Economic Literature*, 51(1):5–62.

[Dufwenberg and Kirchsteiger] Dufwenberg, M. and Kirchsteiger, G. A theory of sequential reciprocity. *Games and Economic Behavior*, 47(2).

[11] Falk, A. and Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, 54(2):293–315.

[12] Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868.

[13] Garcia-Pola, B., Iriberri, N., and Kovářík, J. (2020). Non-equilibrium play in centipede games. *Games and Economic Behavior*, 120:391–433.

[14] Halevy, Y. (2007). Ellsberg revisited: An experimental study. *Econometrica*, 75(2):503–536.

[15] Johnson, E. J., Camerer, C., Sen, S., and Rymon, T. (2002). Detecting failures of backward induction: Monitoring information search in sequential bargaining. *Journal of Economic Theory*, 104(1):16–47.

References III

[16] Knoepfle, D. T., Wang, J. T.-y., and Camerer, C. F. (2009). Studying learning in games using eye-tracking. *Journal of the European Economic Association*, 7(2–3):388–398.

[17] Kuhn, H. W. (1953). Extensive games and the problem of information. In Kuhn, H. W. and Tucker, A. W., editors, *Contributions to the Theory of Games*, volume 2, pages 193–216. Princeton University Press, Princeton.

[18] Levitt, S. D., List, J. A., and Sadoff, S. E. (2011). Checkmate: Exploring backward induction among chess players. *American Economic Review*, 101(2):975–990.

[19] McKelvey, R. D. and Palfrey, T. R. (1992). An experimental study of the centipede game. *Econometrica*, 60(4):803–836.

[20] Nagel, R. (1995). Unraveling in guessing games: An experimental study. *American Economic Review*, 85(5):1313–1326.

[21] Polonio, L. and Coricelli, G. (2019). Testing the level of consistency between choices and beliefs in games using eye-tracking. *Games and Economic Behavior*, 113:566–586.

[22] Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review*, 83(5):1281–1302.

References IV

[23] Selten, R. (1965). Spieltheoretische behandlung eines oligopolmodells mit nachfrageträgheit. *Zeitschrift für die gesamte Staatswissenschaft*, 121:301–324 and 667–689.

[24] Wang, J. T.-y., Spezio, M., and Camerer, C. F. (2010). Pinocchio's pupil: Using eyetracking and pupil dilation to understand truth-telling and deception in sender-receiver games. *American Economic Review*, 100(3):984–1007.

Repetition

Testing Backward Induction

Repetition I — Backward induction and its tests

- Why is it hard to test backward induction directly from observed offers?

Repetition I — Backward induction and its tests

- Why is it hard to test backward induction directly from observed offers?
  - Preferences and reasoning are jointly identified: a wide range of behaviour can be rationalised as SPNE under *some* payoff-interdependent preferences.

Repetition I — Backward induction and its tests

- Why is it hard to test backward induction directly from observed offers?
  - Preferences and reasoning are jointly identified: a wide range of behaviour can be rationalised as SPNE under *some* payoff-interdependent preferences.
- What are subgame consistency and truncation consistency?

Repetition I — Backward induction and its tests

- Why is it hard to test backward induction directly from observed offers?
  - Preferences and reasoning are jointly identified: a wide range of behaviour can be rationalised as SPNE under *some* payoff-interdependent preferences.
- What are subgame consistency and truncation consistency?
  - Subgame consistency: play in a subgame is independent of the subgame's position in a larger game. Example
  - Truncation consistency: replacing a subgame with its equilibrium payoffs does not affect play elsewhere. Example

Repetition I — Backward induction and its tests

- Why is it hard to test backward induction directly from observed offers?
  - Preferences and reasoning are jointly identified: a wide range of behaviour can be rationalised as SPNE under *some* payoff-interdependent preferences.
- What are subgame consistency and truncation consistency?
  - Subgame consistency: play in a subgame is independent of the subgame's position in a larger game. Example
  - Truncation consistency: replacing a subgame with its equilibrium payoffs does not affect play elsewhere. Example
- What are potential reasons for violations of subgame consistency?

Repetition I — Backward induction and its tests

- Why is it hard to test backward induction directly from observed offers?
  - Preferences and reasoning are jointly identified: a wide range of behaviour can be rationalised as SPNE under *some* payoff-interdependent preferences.
- What are subgame consistency and truncation consistency?
  - Subgame consistency: play in a subgame is independent of the subgame's position in a larger game. Example
  - Truncation consistency: replacing a subgame with its equilibrium payoffs does not affect play elsewhere. Example
- What are potential reasons for violations of subgame consistency?
  - Reciprocity; information effects (P2 about P1 through P1's stage-1 offer); selection effects (only certain P2s reach stage 2).

Repetition II — Binmore

- Binmore et al. observe violations of subgame consistency — what exactly and why?

Repetition II — Binmore

- Binmore et al. observe violations of subgame consistency — what exactly and why?
  - Higher offers *and* higher rejection rates in stage 2 of Game III than in Game II.
  - No belief-update story: 2's offer barely depends on 1's rejected demand; 1's rejection rate in stage 2 barely depends on 1's own demand.
  - No selection story: Game-II offers of stage-2 reachers and of all P2s essentially identical.
  - Leading remaining candidate is reciprocity.

Repetition II — Binmore

- Binmore et al. observe violations of subgame consistency — what exactly and why?
  - Higher offers *and* higher rejection rates in stage 2 of Game III than in Game II.
  - No belief-update story: 2's offer barely depends on 1's rejected demand; 1's rejection rate in stage 2 barely depends on 1's own demand.
  - No selection story: Game-II offers of stage-2 reachers and of all P2s essentially identical.
  - Leading remaining candidate is reciprocity.
- Binmore et al. observe violations of truncation consistency — what exactly and why?

Repetition II — Binmore

- Binmore et al. observe violations of subgame consistency — what exactly and why?
  - Higher offers *and* higher rejection rates in stage 2 of Game III than in Game II.
  - No belief-update story: 2's offer barely depends on 1's rejected demand; 1's rejection rate in stage 2 barely depends on 1's own demand.
  - No selection story: Game-II offers of stage-2 reachers and of all P2s essentially identical.
  - Leading remaining candidate is reciprocity.
- Binmore et al. observe violations of truncation consistency — what exactly and why?
  - Player-1 demands respond to rejection payoffs in Game IV (fixed) much more than to the same expected payoffs in Game III (continuation).
  - Candidates: strategic / ambiguity aversion; limited look-ahead.

Repetition III — Johnson et al.

- Johnson et al. use mouse tracking Jump — key hypotheses and observations?

Repetition III — Johnson et al.

- Johnson et al. use mouse tracking Jump — key hypotheses and observations?
  - Limited look-ahead. Level-0 looks only at round 1; level-1 at rounds 1–2; equilibrium players backward-induce.
  - Subjects look most at round 1 and make forward transitions more than backward. Jump
  - Subjects do not backward-induce; behaviour fits limited look-ahead.

Repetition III — Johnson et al.

- Johnson et al. use mouse tracking Jump — key hypotheses and observations?
  - Limited look-ahead. Level-0 looks only at round 1; level-1 at rounds 1–2; equilibrium players backward-induce.
  - Subjects look most at round 1 and make forward transitions more than backward. Jump
  - Subjects do not backward-induce; behaviour fits limited look-ahead.
- Johnson et al. match subjects with robots — why, and what happens?

Repetition III — Johnson et al.

- Johnson et al. use mouse tracking Jump — key hypotheses and observations?
  - Limited look-ahead. Level-0 looks only at round 1; level-1 at rounds 1–2; equilibrium players backward-induce.
  - Subjects look most at round 1 and make forward transitions more than backward. Jump
  - Subjects do not backward-induce; behaviour fits limited look-ahead.

- Johnson et al. match subjects with robots — why, and what happens?
  - Robots eliminate fairness concerns about the opponent.
  - Offers against robots only modestly lower than against humans — far above SPNE.
  - Brief training collapses offers onto SPNE and transfers to a different game.
  - Conclusion: limited look-ahead, not fairness, was binding.

Repetition IV — Synthesis

- Can preference-based models alone explain the BI failures?

Repetition IV — Synthesis

- Can preference-based models alone explain the BI failures?
  - Not fully. Reciprocity rationalises subgame inconsistency; truncation inconsistency needs ambiguity-aversion or limited cognition.

Repetition IV — Synthesis

- Can preference-based models alone explain the BI failures?
  - Not fully. Reciprocity rationalises subgame inconsistency; truncation inconsistency needs ambiguity-aversion or limited cognition.
- Connect the Johnson et al. "level-k" framing to the broader literature.

Repetition IV — Synthesis

- Can preference-based models alone explain the BI failures?
  - Not fully. Reciprocity rationalises subgame inconsistency; truncation inconsistency needs ambiguity-aversion or limited cognition.
- Connect the Johnson et al. "level-k" framing to the broader literature.
  - Same logic as level-k in normal-form games (**?** 20), applied to depth of *look-ahead*; cognitive hierarchy (6) extends.

Repetition IV — Synthesis

- Can preference-based models alone explain the BI failures?
  - Not fully. Reciprocity rationalises subgame inconsistency; truncation inconsistency needs ambiguity-aversion or limited cognition.
- Connect the Johnson et al. "level-k" framing to the broader literature.
  - Same logic as level-k in normal-form games (**?** 20), applied to depth of *look-ahead*; cognitive hierarchy (6) extends.
- When is BI a reliable description of behaviour?

Repetition IV — Synthesis

- Can preference-based models alone explain the BI failures?
  - Not fully. Reciprocity rationalises subgame inconsistency; truncation inconsistency needs ambiguity-aversion or limited cognition.
- Connect the Johnson et al. "level-k" framing to the broader literature.
  - Same logic as level-k in normal-form games (**?** 20), applied to depth of *look-ahead*; cognitive hierarchy (6) extends.
- When is BI a reliable description of behaviour?
  - Short trees, transparent algorithms, experienced/trained subjects, high stakes, no strong fairness motive. Chess players in field centipedes (**?**) are the cleanest positive example.

Repetition IV — Synthesis

- Can preference-based models alone explain the BI failures?
  - Not fully. Reciprocity rationalises subgame inconsistency; truncation inconsistency needs ambiguity-aversion or limited cognition.
- Connect the Johnson et al. "level-k" framing to the broader literature.
  - Same logic as level-k in normal-form games (**?** 20), applied to depth of *look-ahead*; cognitive hierarchy (6) extends.
- When is BI a reliable description of behaviour?
  - Short trees, transparent algorithms, experienced/trained subjects, high stakes, no strong fairness motive. Chess players in field centipedes (**?**) are the cleanest positive example.
- Applied lesson for mechanism design?

Repetition IV — Synthesis

- Can preference-based models alone explain the BI failures?
  - Not fully. Reciprocity rationalises subgame inconsistency; truncation inconsistency needs ambiguity-aversion or limited cognition.
- Connect the Johnson et al. "level-k" framing to the broader literature.
  - Same logic as level-k in normal-form games (**?** 20), applied to depth of *look-ahead*; cognitive hierarchy (6) extends.
- When is BI a reliable description of behaviour?
  - Short trees, transparent algorithms, experienced/trained subjects, high stakes, no strong fairness motive. Chess players in field centipedes (**?**) are the cleanest positive example.
- Applied lesson for mechanism design?
  - Don't assume the SPNE will be played. Design with limited look-ahead in mind, or train participants.