Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Only the Tens Count, but Trumps Decide: Formalising and Benchmarking Dahal Jeet, a Nepali Variant of Mendikot

Ganesh Gautam

ganesh@pcampus.edu.np

Pulchowk Campus https://orcid.org/0000-0002-1926-1492

Raju Kumar Yadav

Pulchowk Campus https://orcid.org/0009-0009-4811-246X

Research Article

Keywords: imperfect-information games, trick-taking card games, determinized search, reinforcement learning, cultural heritage computing, Nepali card games

Posted Date: September 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10955376/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Only the Tens Count, but Trumps Decide: Formalising and Benchmarking Dahal Jeet, a Nepali Variant of Mendikot

Raju Kumar Yadav Ganesh Gautam*

Department of Electronics and Computer Engineering,
IOE, Pulchowk Campus, Tribhuvan University, Lalitpur, Nepal
076bct100.raju@pcampus.edu.np ganesh@pcampus.edu.np

Abstract

Dahal Jeet is a four-player, fixed-partnership, trick-taking card game played daily across the Terai-Madhesh region of Nepal, in which only the four tens score. Its objective is lexicographic and non-additive: a four-valued primary criterion with a trick-count tiebreak that activates on the 2–2 split and decides a third of hands, where the trick-taking games used as benchmarks all score additively. It belongs to the documented Mendikot / Dehla Pakad branch of the Court Piece family, matching Mendikot on 21 of 26 compared features, but the Nepali variant itself has no published rule description: the standard reference's survey of card games played in Nepal lists no game of this branch, and searches of OpenAlex, arXiv, Shodhganga and Nepal Journals Online (NepJOL), in English and Devanagari, return one paper on the entire family. This paper contributes, first, a formal specification of the ruleset drawn from long-term insider participant observation across seven districts and two generations of players; and second, a computational analysis of the game built on that specification. We implement and compare 26 agents spanning hand-written heuristics, metaheuristic tuning, determinized search (PIMC, single- and multiple-observer ISMCTS), and seven families of learning agent, evaluated under duplicate-deal protocols with bootstrap confidence intervals and family-wise error control over 325 pairwise comparisons. Search dominates: the strongest agent is ISMCTS (Elo 1663.2), and no learning method exceeds a tuned heuristic. An exact double-dummy solver provides an endgame optimality bound, showing search converts 0.923 of theoretically won endgames against 0.810 for the best heuristic. We further report a negative result of practical interest: a deliberately minimal state encoding outperforms a rich hand-designed one by +0.0644 win rate, with no seed overlap. Every deterministic heuristic tested is exploitable by a search agent that knows its policy, while random play is not, and a variance decomposition attributes 29–39% of outcome variance to the deal itself. Conditioning on the deal exposes the result the title names: in a game where only the tens score, trump length predicts the outcome far more strongly than the tens dealt, moving the win rate by 0.592 against 0.254. Finally, a human trial with 10 players from 8 districts, 20 hands each, shows why the deal must be controlled before a win rate means anything: the strongest agent's raw rate against them, 0.4800 (95% CI [0.4200, 0.5350], resampling participants), is indistinguishable from chance, yet replaying each participant's own cards with a stand-in agent in their seat gives a paired deficit for the human of -0.0930 (95% CI [-0.1450, -0.0380]). The agent's advantage over people is real and small, and invisible to the unpaired comparison. The protocol is not specific to this game: wherever chance dominates outcome variance, replaying each participant's own realised deal with a stand-in agent recovers a comparison that the unpaired analysis reverses.

Keywords: imperfect-information games | trick-taking card games | determinized search | reinforcement learning | cultural heritage computing | Nepali card games

*Corresponding author: Ganesh Gautam, ganesh@pcampus.edu.np

1 Introduction

A game played every afternoon in Sarlahi has never been written down. Dahal Jeet is a trick-taking game of the Court Piece / Rang family, played by four players in two fixed partnerships, in which *only the four tens score*. The family is documented and we do not claim otherwise: this is a Nepali variant of Mendikot, and §5 sets out the correspondence feature by feature. What is absent is the variant. It is played daily in village gatherings across the Nepal Terai-Madhesh, yet the standard reference's own survey of card games played in Nepal lists no game of this branch, and the peer-reviewed literature contains a single paper on the whole family.

Two contributions follow from that absence. The first is documentation: a formal, testable specification of the rules, with provenance. The second is computational: with a specification and a verified engine, the game becomes a testbed, and the questions that matter for imperfect-information game AI can be asked of it.

The gap. We searched OpenAlex, arXiv and Shodhganga for the game and its documented relatives, and NepJOL, Nepal's own journal index, across eight queries in English and Devanagari. None returns a record of this game, and NepJOL returns nothing on Nepali card games generally. We state the resulting claim precisely: *no record was found in the indexed literature, including Nepal's national journal index*. That is not the same as proving no record exists; much Nepali scholarship appears in print-only venues, and a game may be described under a local name none of our queries guessed.

Beyond the region. The game is local; three of its properties are not, and we separate the two throughout. First, its objective is *lexicographic and non-additive*: a four-valued primary criterion with a thirteen-valued secondary that activates on one configuration of the primary and is otherwise ignored (§3, Equation (6)), where every trick-taking game commonly used as a benchmark scores additively. Second, our principal negative result comes with a mechanism, the scoring density $\rho_{\text{score}} = 4/52$, and that mechanism yields a prediction the game family itself can falsify, since Court Piece scores every trick on a near-identical skeleton. Third, the human-trial protocol is a general remedy for evaluating agents against people in any game where chance dominates outcome variance: on our 200 hands the unpaired and paired analyses support opposite conclusions. §11.1 states all three, and states plainly which are hypotheses rather than measurements.

Contributions. This paper contributes (a) a formal specification of Dahal Jeet with numbered, individually testable rules, a documented provenance across seven districts, and a mathematical model of the ruleset in §3; (b) a reference engine with a behavioural test that no agent can access information its seat has not observed; (c) a comparison of 26 agents under one protocol, with family-wise error control and a retrospective power analysis, in which every reported estimator is defined in §7 and 30 of 30 of the resulting identities are verified against the stored artifacts; (d) an exact double-dummy endgame solver giving an optimality bound; (e) a paired-replay protocol for human evaluation that removes deal luck within each pair, together with the demonstration that the unpaired comparison on the same hands is not merely weaker but misleading; and (f) three negative or corrective results reported in full: reward shaping does not separate under PPO; a minimal encoding beats a rich one; and every deterministic heuristic tested is exploitable by a search that knows its policy, while random play is not.

2 The game

Scope. This paper documents the *single-hand* (immediate capture) variant, which is what the observed player communities play. A deferred-capture variant exists in the wider family and is out of scope.

Seats 1 and 3 play against seats 2 and 4. Each player receives thirteen cards, dealt in batches of five, four and four. Nobody chooses trump. Once the first five are out, the seat after the dealer draws one of its own five at random without looking, and that card's suit is trump for the rest of the hand. All four players see the card. It then goes back into the hand it came from, where it stays playable like any other.

Following suit is compulsory for a player who can. A player who cannot may play anything at all, and is never obliged to trump. The trick falls to the highest trump in it, or, where nobody trumped, to the highest card of the suit led. Its winner takes the cards there and then.

Forty-eight of the fifty-two cards are worth nothing. Three tens or more takes the hand. At two each the tens stop deciding anything and trick count settles it, which it always can, since thirteen tricks cannot split evenly. All four tens is a *coat*; all four tens together with all thirteen tricks is a *double coat*. The losing side deals next. A coat pushes that deal past the seat it would otherwise fall to and on to the loser's partner; a double coat cancels the push.

Family placement. Systematic rule-matching against retrieved primary sources places Dahal Jeet inside the Court Piece / Rang family (McLeod, n.d.c), in its tens-counting branch, closest to Mendikot (McLeod, n.d.f) at 21 of 26 compared features (Table 2), with the double-hand variant reproducing Dehla Pakad's deferred capture (McLeod, n.d.d). The local term *coat* is the South Asian *kot*. §5 sets out the correspondence in full. The dealer rotation and the random trump draw are both attested in the family and are not distinctive to this variant. One rule is unattested in any source we located: that a double coat cancels the coat's dealer-skip.

3 Formal model

The prose specification of §2 is what the player communities describe. This section states the same rules as mathematical objects, so that the quantities reported later have definitions rather than descriptions. Every equation here is implemented in the released engine, and the file and function implementing each one is named where it is not obvious.

Symbol	Meaning
$N = \{1, 2, 3, 4\}$	seats, in turn order
$T_0 = \{1, 3\}$, $T_1 = \{2, 4\}$	the two fixed partnerships
$\mathcal{C}$, $\lvert\mathcal{C}\rvert = 52$	the deck; suit(c), rank(c)
$\mathcal{T} \subset \mathcal{C}$, $\lvert\mathcal{T}\rvert = 4$	the tens: the scoring cards
H_i	the cards seat i still holds
t	the trump suit, fixed once and public
ℓ	the suit led in the current trick
τ	the current trick, a set of (seat, card) pairs
m_k	tens captured by team k; $m_0 + m_1 = 4$
k_k	tricks captured by team k; $k_0 + k_1 = 13$
$u(z) \in \{0, 1\}$	the hand's value to T_0
$\mathcal{I}_i$	the information sets of seat i
V_i	suits seat i is provably void in
d	the dealer's seat
n	number of deals in an evaluation
$r \in \{0, 1, 2, 3\}$	rotation index under duplicate dealing
$\hat{w}$	estimated win rate
ρ	intra-deal correlation
θ, c, s	endgame: won share, conversion, steal

Table 1: Notation used throughout. Seats are numbered 1–4 here and 0–3 in the code.

The game object. Dahal Jeet is a finite extensive-form game with imperfect information, chance moves and two teams:

$$G = \big(N, \{T_0, T_1\}, \mathcal{C}, \mathcal{H}, \mathcal{Z}, \varrho, \{\mathcal{I}_i\}_{i \in N}, u\big), \tag{1}$$

where $\mathcal{H}$ is the set of histories, $\mathcal{Z} \subset \mathcal{H}$ the terminal ones, ϱ the chance measure and u the payoff of (6). Partnerships are fixed and known, so the teams are part of the game rather than a coalition structure to be discovered.

Two chance nodes, not one. The deal is an ordered partition of the deck into four labelled hands of thirteen, and the trump suit is a *second* chance move: the seat after the dealer draws uniformly from its own first five cards, without looking. Writing $H_{d+1}^{(1:5)}$ for those five cards,

$$|\mathcal{D}| = \frac{52!}{(13!)^4} = 5.364 \times 10^{28}, \qquad t \sim \mathrm{Unif}\big\{\mathrm{suit}(c) : c \in H_{d+1}^{(1:5)}\big\}. \tag{2}$$

The consequence is worth stating because it separates this family from most studied trick-taking games: trump is *drawn*, never chosen or bid, so no part of the bidding literature applies and the trump suit carries no information about anyone's hand beyond the one card shown.

Legality. Following suit is compulsory when possible, and discretionary otherwise; there is no obligation to trump when void (`hand.py`, `legal_moves`):

$$L(H_i, \ell) = \begin{cases} \{c \in H_i : \mathrm{suit}(c) = \ell\} & \text{if that set is non-empty,} \\ H_i & \text{otherwise.} \end{cases} \tag{3}$$

The second branch is what generates the paper's central piece of hidden information. A seat that takes it is provably void thereafter,

$$V_i \;\leftarrow\; V_i \cup \{\ell\} \quad \text{whenever seat } i \text{ plays } c \text{ with } \mathrm{suit}(c) \neq \ell, \tag{4}$$

and V_i is public: every seat at the table can draw the same inference. (4) is used three times below, to count information sets exactly (22), to constrain determinization, and to measure how much unconstrained sampling wastes (23).

Trick resolution. A trick is won by the highest trump, or failing that the highest card of the led suit. Both cases collapse into one lexicographic argmax by scoring trumps above the whole rank order:

$$\varphi(c) = \begin{cases} 52 \cdot \mathbf{1}[\mathrm{suit}(c) = t] + \mathrm{rank}(c) & \text{if } \mathrm{suit}(c) \in \{t, \ell\}, \\ -\infty & \text{otherwise,} \end{cases} \qquad w(\tau) = \operatorname*{argmax}_{(i,c) \in \tau} \varphi(c). \tag{5}$$

Cards neither of the led suit nor of trump can never win, which is why φ sends them to $-\infty$ rather than to a low finite value.

The objective. Only the four tens score, and a 2–2 split is decided on trick count. With m_0 tens and k_0 tricks to T_0:

$$u(z) = \mathbf{1}[m_0 \geq 3] \;+\; \mathbf{1}[m_0 = 2] \cdot \mathbf{1}[k_0 \geq 7]. \tag{6}$$

Remark 1. (6) is *lexicographic and non-additive*: a four-valued primary criterion, with a thirteen-valued secondary that activates on exactly the 2–2 split and is otherwise ignored entirely. The trick-taking games that serve as standard benchmarks (bridge, hearts, skat, spades) all score additively, so an agent can trade a small loss on one dimension against a larger gain on another. Here it cannot: no number of tricks compensates for a third ten, and no number of tens is worth anything once the split is 2–2. §11.1 takes up what follows from that.

Since there are thirteen tricks, $k_0 \neq k_1$ always, so (6) admits no draw. The two named results are

$$\mathrm{coat}(z) = \mathbf{1}[m_w = 4], \qquad \mathrm{double}(z) = \mathbf{1}[m_w = 4] \cdot \mathbf{1}[k_w = 13], \tag{7}$$

for the winning team w; a double coat is therefore a coat, which is why R4 below can cancel R3 without ambiguity.

Dealer rotation. The deal passes within the losing side, a coat skips to the loser's partner, and a double coat cancels that skip (`session.py`, `next_dealer`). Writing $\pi(i)$ for i's partner and δ_2 for the first seat of the losing team $T_{\bar{w}}$ in turn order from d,

$$\delta\big(d, \bar{w}, \mathrm{coat}, \mathrm{double}\big) = \begin{cases} \pi(\delta_2) & \text{if coat} = 1 \text{ and double} = 0 \quad \text{(R3)}, \\ \delta_2 & \text{otherwise} \quad \text{(R2, or R4 cancelling R3)}. \end{cases} \tag{8}$$

R4 is the one rule we could not find attested in any documented source (§5).

The game has no terminal state. Play continues indefinitely and what accumulates is a social tally, not a target score. Formally the session is a Markov chain on (d, tally) whose transition is (8), and it has *no absorbing state*: for every reachable state there is a successor. Every episode boundary in this paper, one hand, is therefore a modelling artifact imposed by the experiment, not a rule of the game, and we mark it as such rather than presenting thirteen tricks as the natural horizon.

Scoring density. The property that drives the paper's principal result can be stated as one ratio. Writing ρ_{score} for the fraction of the deck that scores and $E_{\max}$ for the largest number of credit-bearing events a hand can contain,

$$\rho_{\mathrm{score}} = \frac{|\mathcal{T}|}{|\mathcal{C}|} = \frac{4}{52} \approx 0.0769, \qquad E_{\max} = |\mathcal{T}| = 4 \text{ over } 52 \text{ plies}. \tag{9}$$

Forty-eight of fifty-two cards are worth nothing on their own, and a full hand of thirteen tricks yields at most four events that bear on (6). §11 argues that this is why search beats learning here, and §11.1 turns ρ_{score} into a prediction that can be tested on other members of the family.

4 Provenance and method

The specification is a formalisation of the author's own long-term participation, an insider or emic stance that is standard in ethnography and folklore studies. It rests on three cohorts: a college peer group of approximately twenty players drawn from six districts; a documented cohort aged 25–35; and village elders aged 50–80 who play daily after 2pm, with four to five separate groups in a single village and a rotating gathering of fifteen to sixteen at the author's home. All report the same ruleset across seven districts. Agreement across seven districts and two generations is what licenses describing this as a regional variant rather than one village's house rules.

Positionality and declared limits. Insider status provides access and fluency, and is also a bias: familiarity can render local variation invisible. The record is retrospective; there are no contemporaneous field notes; no systematic instrument was administered at the time; and the players were not research participants, because at the time it was not research. Attribution is at district level only, as the players prefer anonymity.

Against that, the documented cohort is the confirmatory record. Those players are named, consenting and contactable, and were taken through the specification rule by rule, including R4, the one rule we could not find attested in any documented source. Their identities are held with the consent register and go to editors or reviewers on request.

5 Related work

The game family, and a literature that does not contain it. Dahal Jeet is not a new game type, and this paper does not claim one. It belongs to the Court Piece / Rang family (McLeod, n.d.c) and, within it, to the branch in which only the tens score. Two games in that branch are documented: Mendikot (McLeod, n.d.f), in which tricks are captured as they are won, and Dehla Pakad (McLeod, n.d.d), in which cards accumulate in the centre until one player wins two consecutive tricks. Our single-hand variant matches Mendikot on 21 of 26 compared features, and our double-hand variant reproduces Dehla Pakad's deferred capture exactly. Table 2 gives the comparison in full. The 5 features that differ are the three trump-determination options Mendikot offers and this variant does not, the trigger for the dealer skip, and the double-coat cancellation.

Vocabulary carries the correspondence further than the mechanics do. Players in the Terai say *coat*, and the word is the South Asian *kot*, which the Mendikot page glosses as a win in which the opponents score nothing (McLeod, n.d.f). That same page traces *mendi* to a Gujarati word for zero, the ten being the only card that contains one. The Terai name for the deferred-capture game, "double hand", turns out to be a calque of a name Pagat already records: Double Sir, or Double Sar, where *sir* glosses as trick (McLeod, n.d.c). Borrowed words are evidence of contact between communities, so we read all three as transmission rather than as coincidence.

Other features look distinctive until one checks the sources. Dealer rotation appears on the Court Piece page, and so does drawing the trump indicator blind from the first player's own five cards (McLeod, n.d.c), so neither is ours to claim. What is unattested is narrow, and we state it narrowly: a double coat cancelling the coat's dealer-skip is unattested in any source we located, as is the pairing of immediate capture with a four-tens-triggered skip, since Mendikot ties that skip to a whitewash and Dehla Pakad pairs the four tens with deferred capture. Wikipedia's Court Piece article (Wikipedia contributors, n.d.) carries a nearby rule but is unsourced in places and we use it as a lead rather than as authority.

The gap this paper fills is therefore regional and evidentiary rather than taxonomic. Pagat's own survey of card games played in Nepal lists Marriage, Call Bridge, Jhyap/Dhumbal, 29 and 3-2-5, and no game of the tens-collecting type (McLeod, n.d.g,n,n,n). No Nepali, Maithili or Bhojpuri source describes this variant, and no peer-reviewed source describes the family's Nepali presence at all.

Computational work on this family. There is one prior study. Juthani (2024) applies prompt engineering to a large language model playing Mendikot. It is the only computational treatment of any game in this branch that we could locate, it reports no reproducible protocol, duplicate-deal control or confidence interval, and it is uncited. We therefore cannot position our agent results against a prior baseline on this game, which is itself part of the motivation for building the specification and the engine first.

Search in imperfect-information card games. Determinized search is the methodological backbone of this study. Ginsberg's GIB established perfect-information Monte Carlo for bridge (Ginsberg, 2001); Cowling, Powley and Whitehouse introduced information-set MCTS, including the multiple-observer form we implement (Cowling et al., 2012a), and later measured ensemble determinization in Magic: The Gathering (Cowling et al., 2012b). Solinas and colleagues combine search with supervised learning in trick-taking games (Solinas et al., 2019). On the partnership side, Amit and Markovitch model partner as well as opponent in bridge bidding (Amit and Markovitch, 2006), and Cohensius and colleagues study bidding in Spades (Cohensius et al., 2020), the closest published partnership trick-taking AI work to ours.

Learning agents. The learning families we compare are standard: deep Q-networks (Mnih et al., 2015) with the double (van Hasselt et al., 2015) and duelling (Wang et al., 2016) refinements, PPO (Schulman et al., 2017), A2C (Mnih et al., 2016), MAPPO under centralised training

Feature	Dahal Jeet	Mendikot	Same
Number of players	four	four	yes
Partnership structure	two fixed pairs, partners opposite	two fixed pairs, partners opposite	yes
Deck	standard 52 cards	standard 52 cards	yes
Rank order	A high down to 2	A high down to 2	yes
Cards per player	thirteen	thirteen	yes
Deal pattern	five, then four, then four	five, then the remainder in fours	yes
Turn rotation	one seat at a time in a fixed direction	one seat at a time in a fixed direction	yes
Trump fixed by a random draw	yes, unseen, from the drawer's own first five	yes, drawn by the player to the dealer's right	yes
Seat that fixes trump	the seat after the dealer	the player to the dealer's right	yes
Trump suit public once fixed	yes	yes, under the random-draw method	yes
Trump card returns to hand and stays playable	yes	not stated in the source	**no**
Hidden-trump option (band hukum)	not played	yes	**no**
Trump-on-first-void option (cut hukum)	not played	yes	**no**
Who leads the first trick	the seat that drew the trump	the player to the dealer's right	yes
Following suit compulsory	yes	yes	yes
No obligation to trump when void	yes, any card may be played	yes, any card may be played	yes
Trick winner	highest trump, else highest of the led suit	highest trump, else highest of the led suit	yes
Capture timing	immediate, every trick	immediate, every trick	yes
Scoring cards	the four tens only	the four tens only	yes
Hand won by tens majority	three or more tens	three or four tens	yes
Tiebreak at two tens each	decided on trick count	the side with seven or more tricks	yes
All four tens is a named result	coat	mendikot	yes
All thirteen tricks is a named result	double coat	52-card mendikot, or whitewash	yes
Next dealer comes from the losing side	yes	yes	yes
A special result skips the deal to the loser's partner	yes, triggered by all four tens	yes, triggered by a whitewash	**no**
A double special result cancels that skip	yes, a double coat cancels it	no such rule	**no**

Table 2: Feature-by-feature comparison against Mendikot, the closest documented game: 21 of 26 features agree. Mendikot is taken from Pagat (McLeod, n.d.f); Dahal Jeet from the specification in §2, as implemented and tested in the released engine.

Feature granularity is a judgement of ours, and splitting or merging rows would change the denominator, so the full list is given rather than the ratio alone.

with decentralised execution (Bayen et al., 2022), policy distillation (Hinton et al., 2015), and the self-play and regret-minimisation line running from CFR (Zinkevich et al., 2007) through NFSP (Heinrich and Silver, 2016) to Deep CFR (Brown et al., 2019). Reward shaping follows the potential-based formulation (Ng et al., 1999), and RLCard (Zha et al., 2020) is the closest existing card-game RL toolkit, though it does not include this family. Self-play from scratch in the AlphaGo Zero sense (Silver et al., 2017) is the ambition these methods inherit; our negative result is that on this game none of them reaches a tuned heuristic.

Traditional games as computational subjects. The closest work on a traditional Nepali card game is Malla's study of Dhumbal (Malla, 2026), which compares heuristic, search and learning agents and reports that a hand-written heuristic outperforms both ISMCTS and reinforcement learning. We note the parallel because our headline result has the same shape on a different game, reached independently. The two studies share no code, no data and no game: Dhumbal is a draw-and-discard game with published rule descriptions, whereas Dahal Jeet is a fixed-partnership trick-taking game with none, which is why the present work begins with documentation rather than with agents. Beyond cards, Nepali board games have received similar treatment in work on Baghchal (Raut et al., 2026; Agarwal and Iida, 2018), and the wider question of how much of a traditional game is skill rather than chance has been quantified for South Asian games including Teen Patti (Banerjee et al., 2025), though not for this family. The framing of such work as cultural-heritage preservation, rather than as benchmark construction alone, has its own literature (Anderson et al., 2010).

6 Engine and evaluation protocol

Each numbered rule has a corresponding test. A separate behavioural test verifies the absence of information leakage: redistributing the hidden cards among opponents, consistently with everything the acting seat has observed, must not change the card the agent plays.

Duplicate dealing. Evaluating on independently shuffled hands makes deal luck a large variance component. Every deal is therefore played four times with the seat assignment rotated, so both agents hold identical cards from every seat and the comparison is paired. That is also what makes the deal, not the hand, the independent unit throughout §7.

Contrary to a common assumption, the trump-drawing seat is *not* structurally advantaged. Across 4,800 self-play hands, the side that drew the trump won 0.5008 of the 2,400 hands in which it did (95% CI [0.4808, 0.5208]), and the seat term in a variance decomposition is 0.000. Rotation is worthwhile because it makes the comparison paired, not because it removes an advantage that exists.

Reproducibility. A refactor and a performance optimisation were applied after results already existed. Table 3 verifies that neither altered any result.

Check	Scope	Result	max $\lvert\Delta\rvert$
Buffer optimisation, all configurations	24 configs	bit-identical	0.0e+00
Stored checkpoint reproduced (120,000 episodes)	rl_dqn_potential_s0.pt	bit-identical	0.0e+00

Table 3: A refactor and a performance optimisation applied mid-project altered no result. In the second row, a checkpoint trained by the original code is reproduced exactly by the current code, which is what licenses reusing earlier artifacts.

In the second row a checkpoint trained by the original code, before both changes, is reproduced bit-for-bit by the current code. Earlier artifacts are therefore both valid and regenerable, which is what licenses reusing them rather than discarding the work.

7 Estimands and estimators

Every headline quantity in §10 is an estimate of something, and this section says what. The estimators are stated in the form the code implements, so that a reader can check the arithmetic rather than take the number on trust.

The identities asserted here are not left as prose. Wherever one is checkable by arithmetic over the stored artifacts, `verify_equations.py` recomputes it, re-running no experiment and reading only files already on disk: 30 of 30 checks pass, and the run is itself an artifact.

The duplicate-deal estimator. Each deal is played four times with the seat assignment rotated. Writing $Y_{jr} \in \{0,1\}$ for whether team A won deal j under rotation r, and $D_j = \sum_{r=0}^{3} Y_{jr} \in \{0,\dots,4\}$ for the per-deal win count,

$$\hat{w}_A = \frac{1}{4n}\sum_{j=1}^{n} D_j = \frac{1}{n}\sum_{j=1}^{n}\frac{D_j}{4}. \tag{10}$$

The deal j, not the hand, is the independent unit: the four rotations share a deal and are therefore correlated, a fact that returns in (17) and decides the intransitivity result.

The per-pairing test. All 325 pairwise comparisons use a two-sided sign-flip permutation test on the per-deal margin, which is exact for this design rather than assuming normality. With statistic

$$T = \frac{1}{n}\sum_{j=1}^{n}\big(D_j - 2\big), \qquad T^*(\varepsilon) = \frac{1}{n}\sum_{j=1}^{n}\varepsilon_j\big(D_j - 2\big), \quad \varepsilon_j \overset{\text{iid}}{\sim} \text{Unif}\{-1,+1\}, \tag{11}$$

the p-value over $B = 20{,}000$ draws is

$$\hat{p} = \frac{1 + \#\{b : |T_b^*| \geq |T|\}}{B+1}. \tag{12}$$

The validity argument is short and is *earned by the duplicate design.* Under the null that the two agents are equally strong, relabelling them maps $D_j \mapsto 4 - D_j$, so $D_j - 2$ is symmetric about zero and sign-flipping is a valid randomisation. That symmetry holds only because both agents played the identical cards from every seat; on independently shuffled hands it would not, and (12) would not be exact.

Multiplicity. With $m = 325$ simultaneous tests, $\alpha m \approx 16.2$ pairings would appear significant by chance. Both corrections are reported, because they answer different questions. For raw p-values sorted $p_{(1)} \leq \cdots \leq p_{(m)}$,

$$\tilde{p}^{\text{Holm}}_{(k)} = \max_{j \leq k}\ \min\big\{1,\ (m-j+1)\,p_{(j)}\big\}, \qquad \tilde{p}^{\text{BH}}_{(k)} = \min_{j \geq k}\ \min\Big\{1,\ \tfrac{m}{j}\,p_{(j)}\Big\}. \tag{13}$$

The running maximum in Holm and the running minimum in BH are what make both sequences monotone in the raw ordering, so a pairing can never be declared significant while a pairing with a smaller p-value is not. Holm controls the family-wise error rate and BH the false-discovery rate, so $\tilde{p}^{\,\text{Holm}} \geq \tilde{p}^{\,\text{BH}}$ pointwise; this is why the surviving counts satisfy $208 \leq 260$ rather than the reverse.

Interval estimates. All confidence intervals are percentile bootstrap intervals that resample *deals.* Drawing indices $j_1^*, \dots, j_n^*$ uniformly with replacement and setting $\hat{w}^* = \frac{1}{4n}\sum_l D_{j_l^*}$ over $B = 2{,}000$ replicates, the interval is $\big[\hat{w}^*_{(0.025)}, \hat{w}^*_{(0.975)}\big]$. Resampling deals rather than hands keeps the four rotations of a deal together, and keeps the correlation between pairings intact, since all pairings share one deal sequence.

Ratings. Pairwise win rates are summarised by a Bradley–Terry fit, solved by the standard MM iteration from $W_i = \sum_j w_{ij}$ wins and n_{ij} games:

$$\Pr(i \succ j) = \frac{p_i}{p_i + p_j}, \qquad p_i \leftarrow W_i \Big/ \sum_{j \neq i} \frac{n_{ij}}{p_i + p_j}, \tag{14}$$

iterated to a fixed point and normalised so that $\prod_i p_i = 1$. The Elo scale is then a relabelling, not a second model:

$$R_i = 1500 + 400 \log_{10} p_i - \frac{1}{|A|} \sum_{j \in A} 400 \log_{10} p_j, \qquad \text{so} \qquad R_i - R_j = 400 \log_{10} \frac{p_i}{p_j}. \tag{15}$$

The right-hand identity is the reason a 400-point gap means ten-to-one odds, and it is one of the identities `verify_equations.py` recomputes. Ratings are relative to this pool.

Transitivity. A rating assumes a single strength ordering exists. Whether one does is measurable. Let $Y_{ij} = w_{ij} - \frac{1}{2}$ be the margin matrix. On a complete comparison graph the least-squares transitive potential has a closed form, because the graph Laplacian is $L = |A|I - J$, so the least-squares solution of $Ls = \operatorname{div} Y$ is the row mean over all $|A|$ entries:

$$s_i = \frac{1}{|A|} \sum_j Y_{ij}, \qquad \kappa = \frac{\sum_{i \neq j} \big(Y_{ij} - (s_i - s_j)\big)^2}{\sum_{i \neq j} Y_{ij}^2}. \tag{16}$$

κ is the share of pairwise-margin energy that no strength ordering can explain: the cyclic component, with $1 - \kappa$ transitive. The decomposition is exact and needs no solver.

Remark 2. Dividing by $|A| - 1$ instead of $|A|$ inflates the potential by $|A|/(|A| - 1)$ and leaves a residual even when Y is a perfect gradient; for three agents that error alone manufactures $\kappa = 0.25$. The released code pins this case in a self-test.

The null for κ. Sampling noise manufactures apparent cycles, so κ means nothing without a floor. The floor is simulated under a *perfectly transitive* truth, each agent's fitted Bradley–Terry strength, with per-deal counts drawn at the over-dispersion actually present in the data. For exchangeable binary outcomes within a deal,

$$\operatorname{Var}(D) = 4p(1-p)\big[1 + 3\rho\big] \qquad \Longrightarrow \qquad \hat{\rho} = \frac{1}{3}\left(\frac{\widehat{\operatorname{Var}}(D)}{4\hat{p}(1-\hat{p})} - 1\right), \tag{17}$$

estimated per pairing and pooled by median. The null then draws $D \mid q \sim \operatorname{Bin}(4, q)$ with $q \sim \operatorname{Beta}\big(pc, (1-p)c\big)$ and concentration $c = (1 - \rho)/\rho$.

Remark 3. The choice of null decides the answer, and the direction of the error is provable rather than a matter of taste. Duplicate dealing gives $\hat{\rho} = 0.097 > 0$, so by (17) the binomial variance $4p(1-p)$ understates the truth by the factor $1 + 3\hat{\rho}$. A binomial null therefore understates the noise, sets the bar too low, and is *anti-conservative* for a claim of intransitivity, the wrong direction for this claim. On this same data it would have reported a spuriously significant result.

Retrospective power. Whether n deals sufficed is a measurement, not a convention. With σ_D the observed per-deal standard deviation, the deals needed to detect a win-rate shift δ two-sided at level α with power $1 - \beta$ are

$$n_{\min}(\delta) = \left\lceil \left(\frac{(z_{1-\alpha/2} + z_{1-\beta})\,\sigma_D}{\delta} \right)^2 \right\rceil, \qquad \delta_{\min}(n) = \frac{(z_{1-\alpha/2} + z_{1-\beta})\,\sigma_D}{\sqrt{n}}. \tag{18}$$

The second form is used for the human trial, where the sample is fixed and the question is what it could have detected.

Variance decomposition. Duplicate dealing turns the question "how much of this is luck" into a two-way layout with one observation per (deal, rotation) cell. With $\bar{Y}$ the grand mean, $\bar{Y}_{j\cdot}$ the deal means and $\bar{Y}_{\cdot r}$ the rotation means,

$$\underbrace{\sum_{j,r}(Y_{jr}-\bar{Y})^2}_{SS_{\text{total}}} = \underbrace{4\sum_{j}(\bar{Y}_{j\cdot}-\bar{Y})^2}_{SS_{\text{deal}}} + \underbrace{n\sum_{r}(\bar{Y}_{\cdot r}-\bar{Y})^2}_{SS_{\text{seat}}} + SS_{\text{residual}}, \tag{19}$$

and the reported shares are

$$\eta^2_{\text{deal}} = \frac{SS_{\text{deal}}}{SS_{\text{total}}}, \quad \eta^2_{\text{seat}} = \frac{SS_{\text{seat}}}{SS_{\text{total}}}, \quad \eta^2_{\text{res}} = 1-\eta^2_{\text{deal}}-\eta^2_{\text{seat}}. \tag{20}$$

SS_{deal} is the cards, i.e. luck; SS_{seat} is structural seat advantage, which rotation removes; the residual is where the matchup actually played out. With one observation per cell there is no replication, so no deal-by-seat interaction can be separated from the residual, and we do not claim one.

Counting information sets exactly. The number of worlds consistent with what a seat has observed is the quantity determinized search samples from, and it can be computed rather than estimated. Let U be the unseen cards, $c_j = 13 - |\text{played by } j|$ the remaining capacity of opponent j, and

$$E(x) = \big\{ j \in N \setminus \{i\} \ : \ \text{suit}(x) \notin V_j \big\} \tag{21}$$

the eligibility set of card x under the voids of (4). Cards with the same eligibility set are interchangeable for counting purposes, so group U by pattern: with three opponents there are at most $2^3 - 1 = 7$ non-empty patterns. Writing n_π for the number of cards with pattern π, the count follows from a dynamic program over patterns whose state is the remaining capacity vector:

$$N\big(\{\pi\} \cup P;\, \mathbf{c}\big) = \sum_{\substack{(k_j)_{j\in\pi}\,:\,k_j\geq 0 \\ \sum_{j\in\pi} k_j = n_\pi}} \frac{n_\pi!}{\prod_{j\in\pi} k_j!}\, N\big(P;\, \mathbf{c}-\mathbf{k}\big), \qquad N(\emptyset;\mathbf{c}) = \mathbf{1}[\mathbf{c}=\mathbf{0}], \tag{22}$$

and $|\mathcal{I}| = N(\Pi;\mathbf{c})$ over all patterns. This is exact, not sampled, and it is verified against brute-force enumeration on small inputs. The unconstrained special cases are the closed forms already given in (2), together with $|\mathcal{I}_1| = 39!/(13!)^3 = 8.448 \times 10^{16}$ at trick 1, when no void has yet been revealed.

What unconstrained sampling wastes. A determinization drawn without regard to the play history may contradict it. The waste is measured directly rather than inferred from win rates: at each decision, $M = 8$ unconstrained worlds are drawn and tested against the observed voids, giving at trick θ

$$\hat{W}_\theta = 1 - \frac{1}{M|\mathcal{S}_\theta|} \sum_{h\in\mathcal{S}_\theta} \sum_{m=1}^{M} \mathbf{1}\big[\text{world } m \text{ consistent with } V(h)\big], \tag{23}$$

over the decisions $\mathcal{S}_\theta$ observed at that trick. This is an estimate with a sampling error, unlike (22), and is reported as one.

8 The structure of the game

There are 5.364×10^{28} distinct deals (exact). A seat's initial information set contains 8.448×10^{16} possibilities (exact). The mean number of legal moves is 3.880, measured over 156,000 decisions.

Trick	Mean legal moves
1	5.715
2	5.465
3	5.269
4	5.004
5	4.737
6	4.477
7	4.221
8	3.841
9	3.465
10	2.993
11	2.453
12	1.796
13	1.000

Table 4: Branching factor by trick, measured over 156,000 decisions of random play. Follow-suit narrows the choice steadily as the hand proceeds.

Trick	Information-set size	Positions sampled
1	$10^{16.93}$	300
4	$10^{12.23}$	300
7	$10^{7.85}$	300
10	$10^{3.69}$	300
12	$10^{1.31}$	300

Table 5: Exact information-set size from one seat's view, with inferred voids and the publicly shown trump card applied, counted by dynamic programming over eligibility patterns.

Uncertainty collapses by roughly fifteen orders of magnitude across a single hand, which is why determinized search is effective here and why sampling budget matters most early.

Uncertainty collapses fast. Evaluating (22) at successive tricks, with the inferred voids and the shown trump card applied, gives $10^{16.93}$ possible worlds at trick 1. By trick 4 that is $10^{12.23}$, then $10^{7.85}$ at trick 7, $10^{3.69}$ at trick 10 and $10^{1.31}$ at trick 12: fifteen orders of magnitude inside one hand. A fixed sampling budget therefore buys incomparable accuracy at trick 2 and trick 11: adequate coverage late, negligible coverage early.

How much is luck. Because every deal is played from all four seat assignments, (19) applies and the question can be answered rather than argued. For TensThenTricks against Random the cards alone account for 0.391 of the variance in outcomes. Seat assignment accounts for 0.000. What is left, 0.609, is the matchup itself.

Trump length against tens dealt. Conditioning on the deal exposes the asymmetry the title names, and §10.17 reports it: trump length dominates tens dealt as a predictor, in a game where only tens score.

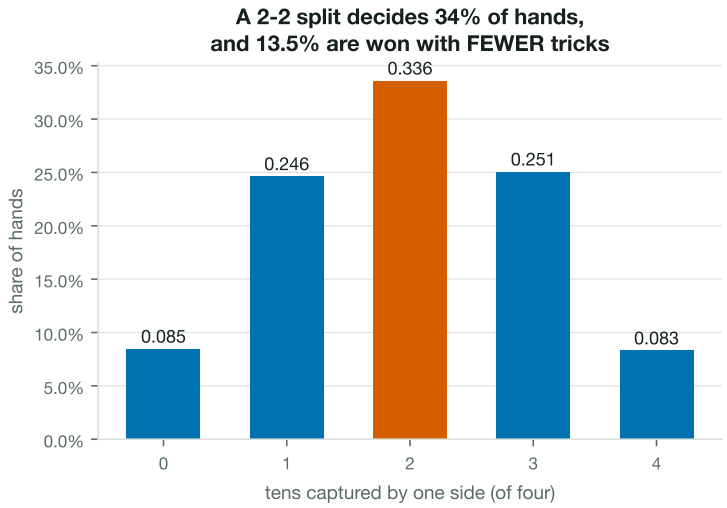


Figure 1: Structural statistics of Dahal Jeet measured over 30,000 random-play hands. Source: `make_figures.py`.

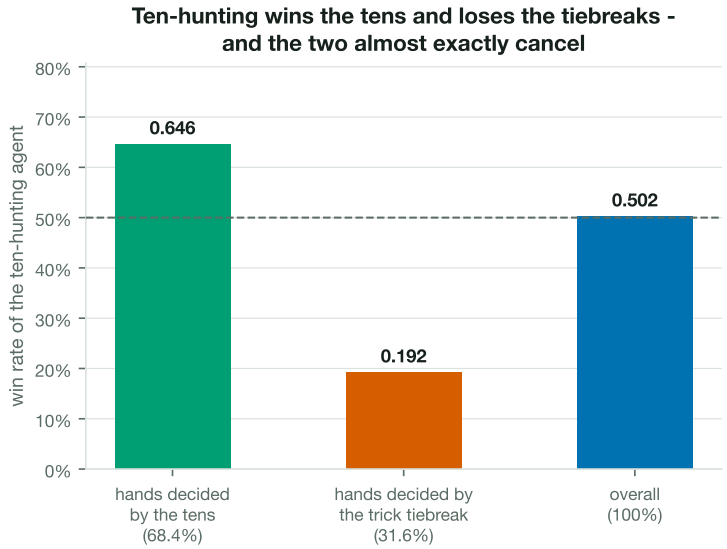


Figure 2: The two objectives pull against each other. Pursuing tens wins most tens-decided hands but loses most 2–2 tiebreaks, and the two effects very nearly cancel. Source: `decomposition.json`.

9 Agents

Heuristics. Seven hand-written agents form a ladder that isolates one idea at a time: random play; greedy trick-taking; partnership awareness (never overtaking one's own partner); ten-awareness; two adaptive policies; and *TensThenTricks*, which pursues tens first and takes tricks as a strict secondary objective. The ladder exists so that any advantage of a complex method can be attributed to something specific rather than to sophistication in general.

Two channels that pull against each other. By (6) a hand is decided on one of two channels: the tens majority, or the trick tiebreak at 2–2. The channels partition the hands, so any agent's overall rate is a mixture,

$$\Pr(\text{win}) = \Pr(\text{win} \mid \text{tens})\Pr(\text{tens}) + \Pr(\text{win} \mid \text{tie})\Pr(\text{tie}), \qquad \Pr(\text{tens}) + \Pr(\text{tie}) = 1, \tag{24}$$

and the two terms move in opposite directions. Measured over 10,000 hands, pure ten-hunting won 0.6460 of tens-decided hands but 0.1922 of tiebreaks; pure trick-taking did the reverse; and

Rank	Heuristic	Mean win rate
1	Tuned	0.6147
2	TensThenTricks	0.5979
3	GreedyTricks	0.5446
4	AdaptiveGreedy	0.5446
5	Balanced	0.5281
6	TenAware	0.5268
7	Adaptive	0.5243
8	PartnerAware	0.5193
9	TenHunter	0.4514
10	TenDefender	0.4363
11	Random	0.3571
12	Lowest	0.3550

Table 6: Hand-written heuristic ladder, 1200 duplicate deals per pairing. Each rung adds one idea to the rung below advantage can be attributed to something specific.

(24) shows why the two sit at 0.5024 head to head rather than one dominating. Figure 2 plots both channels and the mixture, and (24) is one of the identities checked against the artifact.

Metaheuristic tuning. Hand-tuning the balance between the two channels failed, which is itself informative. The balance point is easier to search for than to guess, so a parameterised agent with a numeric genome is tuned offline by hill-climbing against a fixed opponent panel under duplicate-deal evaluation.

Determinized search. PIMC and ISMCTS both sample hypothetical worlds consistent with everything the acting seat has observed. Sampling respects the inferred voids of (4): a seat that failed to follow a suit cannot hold it. Writing $\mathcal{W}(I)$ for the worlds consistent with information set I and σ for the rollout policy, PIMC scores each legal move by the mean payoff of playing it and rolling out:

$$\hat{Q}_{\mathrm{PIMC}}(a) = \frac{1}{W}\sum_{k=1}^{W} u\big(\sigma \mid a, \omega_k\big), \qquad \omega_k \sim \mathrm{Unif}\,\mathcal{W}(I), \tag{25}$$

and plays $\operatorname{argmax}_a \hat{Q}_{\mathrm{PIMC}}(a)$. Each ω_k is solved as though the world were common knowledge, so (25) estimates the mean over worlds of a perfect-information value rather than the value of the information set itself; the two differ, and PIMC cannot value information-gathering as a result.

ISMCTS instead builds one tree keyed by the action sequence from the root, so nodes are information sets, and draws a fresh world per iteration. Selection is UCB1 over the moves legal *in the current world*, with the exploration term normalised by how often each move was *available* rather than by the parent's visit count:

$$a^{\star} = \operatorname*{argmax}_{a \in L(H_i,\ell)} \left[\bar{Q}(a) + c\sqrt{\frac{\log \sum_{a' \in L(H_i,\ell)} n(a')}{n(a)}} \right], \qquad c = 0.7. \tag{26}$$

The availability denominator is the whole difference between (26) and plain MCTS: a move that is legal in few sampled worlds is not penalised for the iterations in which it could not be chosen.

We also implement a multiple-observer variant, and the correction is one subscript. Both forms run (26); they differ only in what a visited node is credited with after the rollout reaches a terminal z:

$$\text{SO:}\quad v(h) \mathrel{+}= u_{T(\mathrm{root})}(z), \qquad \text{MO:}\quad v(h) \mathrel{+}= u_{T(\mathrm{seat}(h))}(z). \tag{27}$$

The single-observer form backs up the *root team's* payoff at every node, including opponents' decision nodes, and then maximises it there, which is to say it searches opponents as though

they were helping the searcher. The multiple-observer variant keeps a separate tree per seat and credits each node with the payoff of the side that actually moves there.

Learning agents. DQN with double and dueling variants; PPO; A2C; MAPPO under centralised training with decentralised execution, which the fixed partnerships make natural; NFSP; Deep CFR; and a policy distilled from ISMCTS. The observation is seat-relative, so a policy learned in one seat transfers to all four.

The four reward schemes. By (9) the honest reward is close to unusable: one signal at the end of a hand, and nothing in between. Rather than pick a shaping function and report whatever it gave, we treat the choice as the experiment. Write $R(z) = +1$ if the learner's team won and -1 otherwise, Δm and Δk for the tens and tricks gained on a transition, and T' for the opposing team. The four schemes are

$$r_{\text{terminal}} = R(z)\,\mathbf{1}[\text{done}], \tag{28}$$

$$r_{\text{ten}} = 0.25\,\Delta m_T - 0.25\,\Delta m_{T'} + R(z)\,\mathbf{1}[\text{done}], \tag{29}$$

$$r_{\text{potential}} = \Phi(s') - \Phi(s) + R(z)\,\mathbf{1}[\text{done}], \quad \Phi(s) = 0.66\,\frac{m_T - m_{T'}}{4} + 0.34\,\frac{k_T - k_{T'}}{13}, \tag{30}$$

$$r_{\text{trick}} = 0.05\,\Delta k_T + R(z)\,\mathbf{1}[\text{done}]. \tag{31}$$

Only (30) has the form $\gamma\Phi(s') - \Phi(s)$ with $\gamma = 1$, so by Ng et al. (1999) only it is guaranteed policy-invariant: it changes credit assignment without moving the optimal policy. (29) and (31) are not of that form and can distort the objective, which is why they are included. The weights 0.66/0.34 inside Φ approximate the measured channel split of (24), which runs from 0.66 to 0.71 on the tens channel across the pairings we decomposed; they are an approximation to that measurement, not the measurement itself.

(31) is a deliberately misaligned control. Trick count is the wrong objective under (6), since Figure 1 shows 13.5% of hands are won by the side taking *fewer* tricks. If the reward ablation separates anything, this arm should therefore lose. §10 reports that under PPO it does not.

10 Results

Every table in this section is generated directly from its result artifact; no value is transcribed by hand.

10.1 Full round-robin

All 26 agents played a complete round-robin of 325 pairings at 300 duplicate deals each, 390,000 hands. Table 7 gives the full ranking and Figure 3 the hierarchy.

Strongest agents. Determinized search takes the top of the table outright. ISMCTS250 leads at 1663.2 and PIMC16 follows at 1643.2, and the gap from there down to the third place is wider than the gap between any other adjacent pair near the top.

Middle of the table. Third place goes to the metaheuristically tuned heuristic, Tuned at 1563.4, ahead of *every* network we trained. MAPPO is the strongest of the learners and the only one that reaches the band the better hand-written heuristics occupy.

Weakest agents. NFSP sits second from bottom, above only Random. We state that plainly rather than leaving it to be read off the table: a method built for imperfect-information games finished below heuristics that fit in a page of code. At the budget we gave it, it did not transfer to this game.

Rank	Agent	Elo	95% CI
1	ISMCTS250	1663.2	[1654.2, 1673.2]
2	PIMC16	1643.2	[1634.4, 1652.4]
3	Tuned	1563.4	[1555.7, 1571.7]
4	AlphaDJ60	1553.2	[1547.8, 1558.5]
5	mappo_potential	1549.2	[1542.4, 1555.1]
6	TensThenTricks	1547.4	[1540.2, 1556.7]
7	PartnerAware	1539.1	[1532.0, 1548.2]
8	distill	1537.0	[1530.3, 1543.3]
9	GreedyTricks	1530.2	[1523.2, 1537.7]
10	ppo_trick_shaped	1519.5	[1512.1, 1526.5]
11	AdaptiveGreedy	1516.0	[1507.9, 1524.0]
12	TenAware	1514.6	[1506.2, 1522.7]
13	Adaptive	1511.7	[1503.3, 1519.2]
14	ppo_potential	1504.5	[1497.5, 1511.3]
15	ppo_terminal	1500.9	[1492.2, 1508.0]
16	ppo_ten_shaped	1490.4	[1483.5, 1497.3]
17	deepcfr	1463.7	[1456.4, 1471.8]
18	double_dueling	1458.9	[1451.8, 1465.5]
19	dqn_potential	1441.2	[1433.5, 1447.7]
20	dqn_terminal	1440.8	[1434.0, 1447.6]
21	dqn_ten_shaped	1439.6	[1431.7, 1446.2]
22	double	1438.0	[1430.2, 1444.2]
23	a2c	1437.4	[1428.8, 1446.0]
24	dqn_trick_shaped	1436.7	[1428.6, 1444.9]
25	nfsp	1387.0	[1378.1, 1395.2]
26	Random	1373.0	[1366.4, 1381.6]

Table 7: Full round-robin ranking, all 26 agents, 300 duplicate deals per pairing. Bradley–Terry ratings on an Elo scale; intervals bootstrapped over deals. Ratings are relative to this pool only.

Search against learning. At the budgets used here, no learning method reaches the tuned heuristic, and determinized search is stronger than both. The gap between the tuned heuristic and every network is the study's principal negative result, and §11 takes up why.

10.2 Statistical control

Of the 325 pairings, 263 are nominally significant, against the 16.2 expected by chance alone at $\alpha = 0.05$; 208 survive Holm–Bonferroni and 260 survive Benjamini–Hochberg. We report corrected counts throughout.

Quantity	of 325 pairings
Nominally significant ($\alpha = 0.05$)	263
Expected false positives if uncorrected	16.2
Survive Holm–Bonferroni	208
Survive Benjamini–Hochberg	260

Table 8: Family-wise error control over the full round-robin. The family is declared as all pairings, before analysis.

10.3 Retrospective power

Substituting the observed per-deal standard deviation, 0.2778, into (18) puts the detectable difference at 300 deals at roughly 0.05, for 80% power. Resolving 0.02 would take 1514 deals, five times the tournament. Anything close to 0.02 is therefore beyond what this design can resolve,

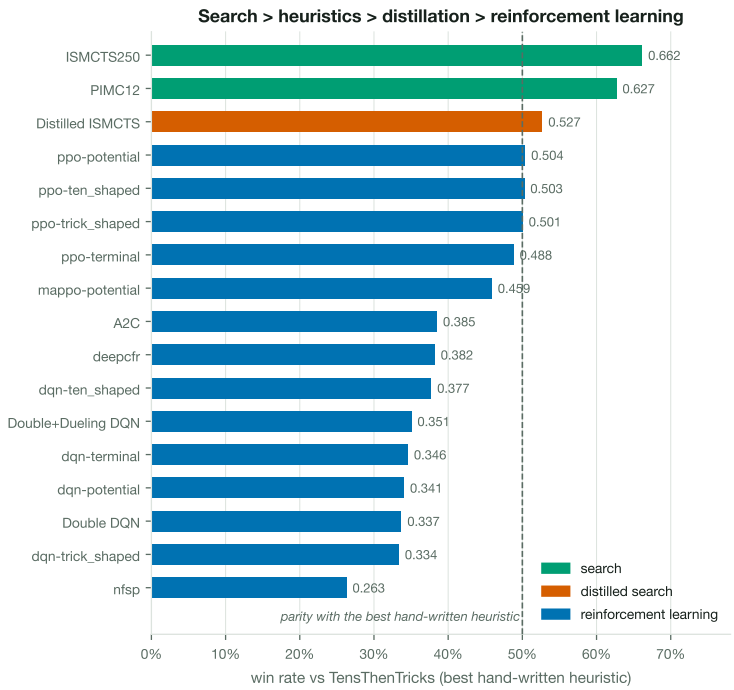


Figure 3: Final ranking across all 325 pairings. Search occupies the top, the tuned heuristic is third, and no learning method exceeds it. Source: `final_tournament.json`.

and we call such differences inconclusive rather than ties. The distinction is not pedantry: it decides how the void-constraint result below has to be read.

10.4 Intransitivity

The cyclic energy of (16) comes to $\kappa = 0.0518$. Its noise floor, simulated under a perfectly transitive Bradley–Terry truth at the measured intra-deal correlation $\rho = 0.097$, is 0.0154 with a 95th percentile of 0.0175, so $p = 0.005$. Counting directly, 66 of the 2,600 triples are three-cycles.

Which null one picks decides that verdict, and Remark 3 gives the reason: at $\rho = 0.097$ a binomial null understates the variance by a factor of $1 + 3\rho$, sets the bar too low, and would have declared this same data significant for the wrong reason. Against the correct floor the finding survives, but it is small. The agents form a mostly clean ladder with a modest cyclic component on top of it.

10.5 Search: strength against cost

Both families gain strength monotonically as the budget rises (Figure 4), and at comparable budgets across a full hand ISMCTS is the stronger of the two.

10.6 Void-constrained determinization

Sampling worlds uniformly ignores the play history, and most such worlds contradict it. Table 10 reports $\hat{W}_\theta$ of (23) directly.

Waste rises from 1.4% at trick 1 to 64% by trick 11 under random play, and from 1.1% to 63% under skilled play, so the constraint addresses a real source of error. Whether it pays is a separate question, and the answer is equivocal.

At matched samples the constrained sampler wins at every budget (Supplementary Table S5). At matched wall-clock, which is the operative comparison since constraining costs time, the advantage narrows to 0.02–0.03, inside the inconclusive band of (18), and only one of eight arms survives Holm–Bonferroni. We report one significant arm of eight, not a general result.

Family	Budget	Win rate	95% CI	ms/decision
PIMC	2	0.4771	[0.4396, 0.5125]	1.75
PIMC	4	0.5312	[0.4958, 0.5687]	3.54
PIMC	8	0.6229	[0.5833, 0.6604]	7.14
PIMC	16	0.6708	[0.6312, 0.7104]	14.48
PIMC	32	0.7312	[0.6958, 0.7667]	28.92
ISMCTS	50	0.5958	[0.5563, 0.6354]	9.49
ISMCTS	100	0.6292	[0.5875, 0.6750]	18.93
ISMCTS	200	0.7021	[0.6646, 0.7396]	38.10
ISMCTS	400	0.7063	[0.6667, 0.7438]	76.09
ISMCTS	800	0.7396	[0.7021, 0.7771]	152.53

Table 9: Search strength against measured decision cost, versus the strongest heuristic. Budget is worlds for PIMC and iterations for ISMCTS.

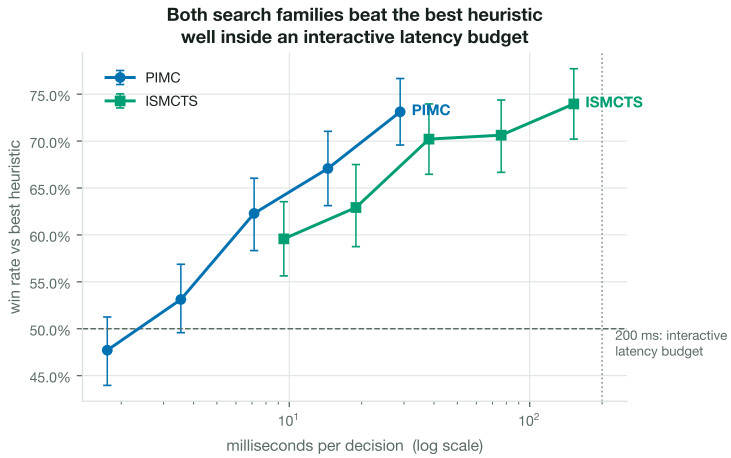


Figure 4: Strength against measured decision cost for both search families. No cross-game reference point is plotted: latency depends on the action space and the rules searched, so a figure from another game is not commensurable. Source: `eval_curve.json`.

Trick	Wasted (random play)	Wasted (skilled play)
1	0.0142	0.0112
2	0.0544	0.0675
3	0.1304	0.1215
4	0.2408	0.2406
5	0.3477	0.3408
6	0.4219	0.4854
7	0.5273	0.5731
8	0.6015	0.6321
9	0.6400	0.6596
10	0.6444	0.6773
11	0.6402	0.6277
12	0.5417	0.5150
13	0.1683	0.2017

Table 10: Fraction of unconstrained determinizations that contradict the observed play. Measured under both random and skilled play, since the rate at which voids are revealed differs.

Family	Budget	Constrained	Matched budget	Unconstrained	Δ	Time error	Paired p
ISMCTS	50	0.5631	58	0.5687	-0.0056	-0.8%	0.7107
ISMCTS	100	0.6181	117	0.6144	+0.0037	-0.0%	0.8257
ISMCTS	200	0.6706	236	0.6325	+0.0381	+0.2%	0.0025
ISMCTS	400	0.6694	473	0.6462	+0.0232	+0.1%	0.0716
PIMC	4	0.5363	4	0.5238	+0.0125	-3.6%	0.3702
PIMC	8	0.6075	8	0.5806	+0.0269	-3.7%	0.0487
PIMC	16	0.6556	16	0.6231	+0.0325	-3.5%	0.0146
PIMC	32	0.7025	33	0.6787	+0.0238	-0.8%	0.0590

Table 11: Void-constraint ablation, matched wall-clock. The unconstrained sampler is actually run at a budget chosen to match the constrained arm's measured time; both arms play the same deals, so the test is paired.

Eight simultaneous tests. Under Holm–Bonferroni only the ISMCTS-200 arm survives at $\alpha = 0.05$.

10.7 Opponent modelling in search

Correcting the single-observer backup of (27) is worth between 0.0124 and 0.0500 win rate depending on the search budget, and wins head to head at the largest budget tested.

Variant	Iterations	Win rate	95% CI	ms/dec
SO-ISMCTS	100	0.6138	[0.5875, 0.6400]	19.93
MO-ISMCTS	100	0.6262	[0.5975, 0.6525]	20.01
SO-ISMCTS	200	0.6450	[0.6175, 0.6725]	39.07
MO-ISMCTS	200	0.6950	[0.6663, 0.7212]	38.88
SO-ISMCTS	400	0.6700	[0.6412, 0.7000]	76.14
MO-ISMCTS	400	0.7050	[0.6750, 0.7338]	76.66

Table 12: Single-observer versus multiple-observer ISMCTS. The single-observer form backs up the root team's payoff at every node, so opponents are searched as though they were helping the searcher.

Iterations	MO win rate vs SO	95% CI	Verdict
100	0.5038	[0.4775, 0.5288]	tie
200	0.5225	[0.4988, 0.5475]	tie
400	0.5375	[0.5138, 0.5625]	MO wins

Table 13: Multiple- against single-observer ISMCTS, head to head on identical deals.

10.8 Distance from optimal

Exact double-dummy solving is affordable only in the endgame: cost grows roughly tenfold per additional trick, and whole-hand solving is not computable within any budget we could justify (Supplementary Table S4). We therefore report an *endgame* bound.

The bound needs three quantities and one identity. Let $V^*(z) \in \{0, 1\}$ be the double-dummy value of a handover position for the agent's team. The solver optimises the win indicator of (6) directly, which is also what makes it tractable, since a boolean value set lets minimax short-circuit. Then

$$\theta = \Pr\big(V^* = 1\big), \qquad c = \Pr\big(\text{win} \mid V^* = 1\big), \qquad s = \Pr\big(\text{win} \mid V^* = 0\big), \tag{32}$$

are the share of positions that are theoretically won, the *conversion* rate on those, and the *steal* rate on the rest. Since V^* partitions the positions, the realised win rate a and the reported gap follow by the law of total probability:

$$a = c\,\theta + s\,(1 - \theta), \qquad \text{gap} = \theta - a. \tag{33}$$

(33) is what makes the two tables below one measurement rather than two, and it defines the gap column: the distance between what the cards were worth and what the agent got. A negative gap means the agent's steals outweighed its conversion losses. Both tables satisfy (33) to the precision at which they are stored, and this is checked against the artifacts rather than asserted.

Agent	Theoretical	Actual	Conversion	Steal	Gap
Random	0.3833	0.2533	0.548	0.070	+0.1300
GreedyTricks	0.4883	0.4450	0.747	0.156	+0.0433
TenAware	0.4533	0.4233	0.713	0.183	+0.0300
Adaptive	0.4533	0.4250	0.721	0.180	+0.0283
TensThenTricks	0.5000	0.5000	0.810	0.190	+0.0000
PIMC:16	0.5300	0.6467	0.931	0.326	-0.1167
ISMCTS:200	0.5083	0.6300	0.902	0.349	-0.1217

Table 14: Endgame optimality with each agent generating its own positions. Conversion rates are not comparable across agents here, because a stronger agent reaches different endgames.

Agent	Theoretical	Actual	Conversion	Steal	Gap
Random	0.5000	0.3683	0.630	0.107	+0.1317
GreedyTricks	0.5000	0.4350	0.713	0.157	+0.0650
TenAware	0.5000	0.4817	0.773	0.190	+0.0183
Adaptive	0.5000	0.4833	0.780	0.187	+0.0167
TensThenTricks	0.5000	0.5000	0.810	0.190	+0.0000
PIMC:16	0.5000	0.6200	0.923	0.317	-0.1200
ISMCTS:200	0.5000	0.5900	0.897	0.283	-0.0900

Table 15: Endgame optimality on a common position set: one fixed policy drives every seat to the handover, so all agents face identical positions. The theoretical rate is identical by construction, which is the protocol's own check. These are the comparable numbers.

The two protocols matter. When each agent generates its own positions, conversion rates are not comparable, because a stronger agent reaches different endgames; its theoretical win rate ranges from 0.383 to 0.530. On a common position set the theoretical rate is 0.5000 for every agent by construction, and the comparison is clean: PIMC converts 0.923 of theoretically won endgames and ISMCTS 0.897, against 0.810 for the best heuristic and 0.630 for random play. Search therefore plays endgames close to optimally; heuristics leave roughly one in five won positions on the table.

10.9 Reward design

Under DQN, ten-shaped rewards win consistently across seeds. Under PPO *nothing separates*: every pairwise difference is small and only two of six are consistent in direction across seeds. The deliberately misaligned control, which rewards trick-taking, the secondary objective, does not underperform under PPO.

The misaligned control does not underperform. A plausible reason, rather than a resolution: a 2–2 split of the tens is decided on trick count, so trick-taking is less misaligned with the true objective than the label suggests.

10.10 All learning methods

The strongest learner is distill at 0.5270 ± 0.0096. Untrained networks of the same architectures score 0.260–0.267, and this is the floor against which trained results should be read: a randomly initialised network is not random play, because the architecture and action masking already impose structure. Reporting "beats Random" without this control would overstate what training achieved. The learning curves of Figure 6 flatten well before the budget is exhausted.

Algo	A	B	Pairs	mean(A−B)	SD	Consistent
dqn	dqn-potential	dqn-ten_shaped	3	-0.0367	0.0133	yes
dqn	dqn-potential	dqn-terminal	3	-0.0053	0.0191	no
dqn	dqn-potential	dqn-trick_shaped	3	+0.0069	0.0041	yes
dqn	dqn-ten_shaped	dqn-terminal	3	+0.0314	0.0204	yes
dqn	dqn-ten_shaped	dqn-trick_shaped	3	+0.0436	0.0143	yes
dqn	dqn-terminal	dqn-trick_shaped	3	+0.0122	0.0232	no
ppo	ppo-potential	ppo-ten_shaped	3	+0.0011	0.0110	no
ppo	ppo-potential	ppo-terminal	3	+0.0155	0.0131	no
ppo	ppo-potential	ppo-trick_shaped	3	+0.0033	0.0147	no
ppo	ppo-ten_shaped	ppo-terminal	3	+0.0144	0.0069	yes
ppo	ppo-ten_shaped	ppo-trick_shaped	3	+0.0022	0.0041	no
ppo	ppo-terminal	ppo-trick_shaped	3	-0.0122	0.0096	yes

Table 16: Reward schemes compared seed by seed. Pairing on the seed removes the seed's own contribution; 'consistent' means every seed agreed on the direction.

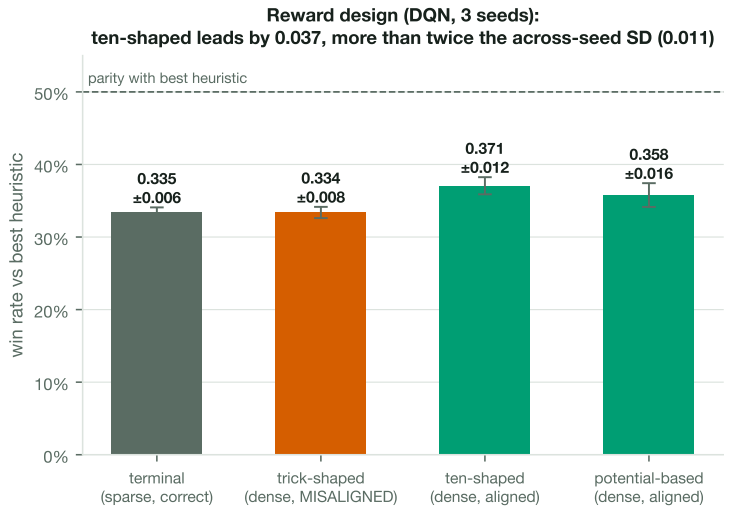


Figure 5: Reward design under DQN, mean ± SD over three seeds. Source: `rl_dqn_*.json`.

Method	Seeds	vs Random	vs best heuristic
distill	3	0.7358 ± 0.0068	0.527 ± 0.0096
ppo-potential	3	0.6921 ± 0.0038	0.5039 ± 0.0086
ppo-ten_shaped	3	0.6853 ± 0.0039	0.5028 ± 0.004
ppo-trick_shaped	3	0.6946 ± 0.0083	0.5006 ± 0.0064
ppo-terminal	3	0.6844 ± 0.0065	0.4883 ± 0.01
mappo-potential	3	0.7187 ± 0.0026	0.4589 ± 0.0186
a2c	3	0.6218 ± 0.0245	0.385 ± 0.0291
deepcfr	3	0.6314 ± 0.0085	0.382 ± 0.013
dqn-ten_shaped	3	0.6103 ± 0.0128	0.3772 ± 0.0193
double_dueling	3	0.6132 ± 0.005	0.3506 ± 0.0109
dqn-terminal	3	0.599 ± 0.0211	0.3458 ± 0.0284
dqn-potential	3	0.614 ± 0.0061	0.3405 ± 0.0094
double	3	0.5936 ± 0.0126	0.337 ± 0.0137
dqn-trick_shaped	3	0.5906 ± 0.0137	0.3336 ± 0.0057
nfsp	3	0.5333 ± 0.0022	0.263 ± 0.0034

Table 17: All learning methods, mean ± standard deviation across seeds.

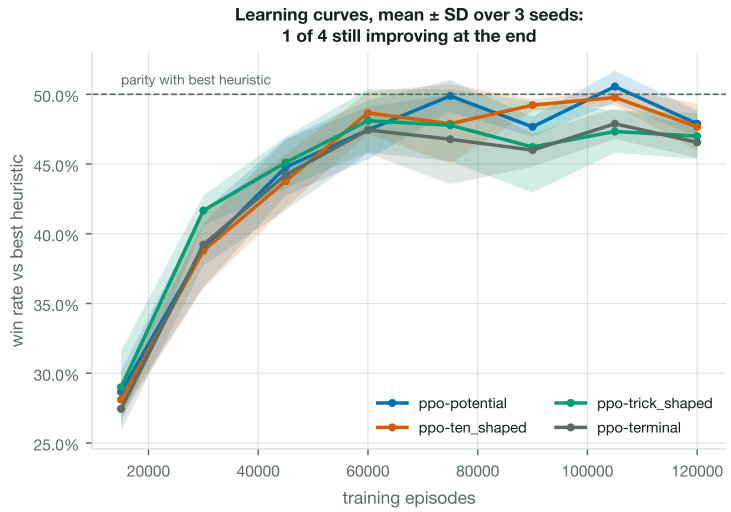


Figure 6: Learning curves, mean ± one standard deviation across seeds. The curves flatten well before the training budget is exhausted, so the shortfall against search is not simply insufficient training. Source: `rl_*_s*.json`.

10.11 Optimisation budget

Matching on episodes, as is conventional, gives DQN and PPO very different optimisation budgets: DQN performs 122× more gradient steps per episode. We therefore matched on wall-clock and again on gradient steps. PPO wins under both, so its advantage is a property of the algorithm on this game rather than of the budget it happened to receive (Supplementary Table S2).

10.12 Generalisation

Every learner is trained against one fixed opponent, which leaves open whether it learned the game or that opponent. Separating the two requires care about what is being differenced. Writing π_o for a policy trained against opponent o and $w(\pi, o)$ for its win rate against o, the measure we report is

$$\sigma_o = w(\pi_o,\, o) \;-\; \frac{1}{|O| - 1} \sum_{o' \neq o} w(\pi_{o'},\, o), \tag{34}$$

in which *both* terms are win rates against the same opponent o, so opponent strength cancels and what remains is the advantage of having trained on that specific opponent. The obvious alternative differences across opponents instead,

$$\text{naive}_o = w(\pi_o,\, o) \;-\; \frac{1}{|B|} \sum_{b \in B} w(\pi_o,\, b), \qquad B = \text{held-out opponents}, \tag{35}$$

and on this data (35) reports +0.185 while (34) reports +0.0323 ± 0.0174. The difference between the two is not a discrepancy: (35) mostly measures that the held-out opponents, which are search agents, are stronger than the training heuristics, which is a fact about B rather than about overfitting. Read through (34), the advantage is small and the agents largely learned the game.

This measure is deliberately constructed to cancel opponent strength: both terms are win rates against the *same* opponent. A naive alternative, win rate against the training opponent minus win rate against held-out opponents, reports +0.185 on the same data, but that compares performance against different opponents and mostly measures that search agents are stronger than heuristics.

Algo	Opponent	Trained on it	Trained elsewhere	Specificity
dqn	Random	0.5690	0.5793	-0.0102
dqn	GreedyTricks	0.3697	0.3280	+0.0418
dqn	TenAware	0.3689	0.3443	+0.0246
dqn	Adaptive	0.3817	0.3497	+0.0320
dqn	TensThenTricks	0.3436	0.3049	+0.0387
ppo	Random	0.7626	0.6994	+0.0632
ppo	GreedyTricks	0.4822	0.4464	+0.0358
ppo	TenAware	0.5364	0.5095	+0.0268
ppo	Adaptive	0.5467	0.5149	+0.0318
ppo	TensThenTricks	0.4995	0.4608	+0.0386

Table 18: Opponent specificity: the advantage from having trained against this specific opponent, with opponent strength cancelled because both terms are measured against the same opponent.

Mean +0.0323 ± 0.0174 over 10 opponents. A naive gap (own minus held-out) would instead report roughly +0.30, but that compares performance against different opponents and mostly measures their relative strength.

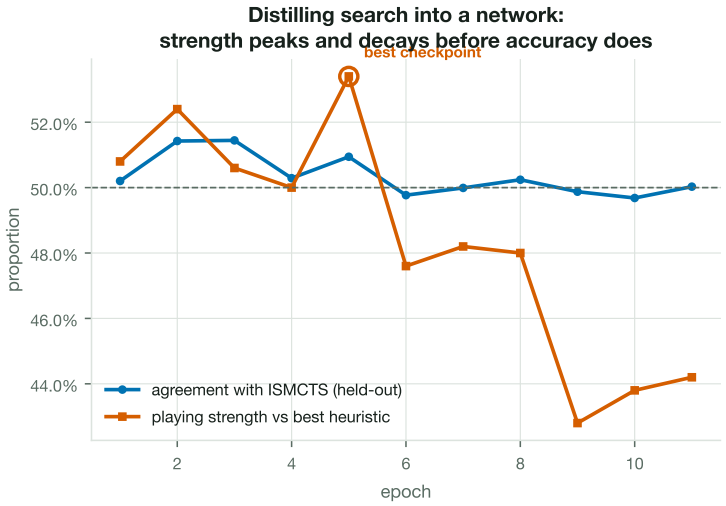


Figure 7: Distillation from ISMCTS. Validation accuracy is flat while playing strength peaks and then decays, which is why checkpoints are selected on win rate rather than on loss. Source: `rl_distill_s*.json`.

10.13 Exploitability

Nash exploitability is only well defined for two-player zero-sum games, and Dahal Jeet is a four-player partnership game, so we measure *best-response* exploitability against an explicit oracle class. With $\mathfrak{B}_T$ the policies trained against target T and nothing else, and $\mathfrak{G}_T$ those that never trained against T,

$$e(T) = \max_{B \in \mathfrak{B}_T} w(B, T) \; - \; \max_{G \in \mathfrak{G}_T} w(G, T). \tag{36}$$

Remark 4. (36) is a *lower bound* on the true best-response gap, and only that: it is exact only if $\mathfrak{B}_T$ contains a genuine best response. An oracle class too weak to beat T makes the first term small and reports $e(T) \leq 0$ whatever T's real exploitability is. This is not hypothetical, and it is not specific to one algorithm: earlier DQN and PPO oracles both finished weaker than the heuristics they were meant to exploit, returning -0.11 to -0.20 on these same targets, and those numbers had to be discarded rather than reported, since a best response that loses to an off-the-shelf heuristic bounds nothing.

Of 5 targets, 4 are significantly exploitable. Random is not (+0.0033, $p = 0.846$), which is the sanity result: there is no pattern in a random policy to exploit, so (36) should return zero, and it does. Every deterministic heuristic is exploitable, including the strongest one (+0.0517).

Knowing the opponent is worth between four and seven percentage points to an otherwise identical search.

Target	Best response	Generic search	Exploitability	Paired p
GreedyTricks	0.7525	0.6817	+0.0708	<0.001
Adaptive	0.7683	0.7042	+0.0642	<0.001
TensThenTricks	0.7075	0.6558	+0.0517	<0.001
TenAware	0.7508	0.7108	+0.0400	0.0093
Random	0.8492	0.8458	+0.0033	0.8465

Table 19: Exploitability: a search that simulates the target's true policy for the opponent seats, against the same search at the same budget with no opponent model. Identical algorithm in both arms, paired on identical deals.

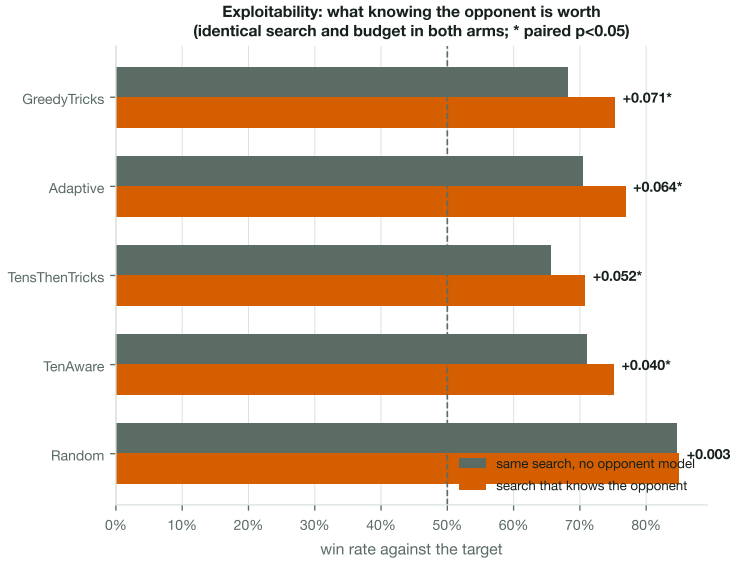


Figure 8: Exploitability. Both arms are the same search at the same budget; only the opponent model differs. Source: `exploitability_search.json`.

10.14 State encoding

No single block's removal changes performance materially; every one-block ablation lands within noise of the full encoding. Yet removing almost all of it, leaving only the agent's own hand and the current trick, *improves* performance by +0.0644, and the seeds of the two conditions do not overlap.

Blocks zeroed	Mean	SD	Seeds	Per-seed
minimal	0.4273	0.0262	3	0.4081, 0.4094, 0.4644
no_scalars	0.3671	0.0091	3	0.3550, 0.3694, 0.3769
no_voids	0.3642	0.0112	3	0.3513, 0.3625, 0.3787
no_ten_status	0.3634	0.0217	3	0.3394, 0.3588, 0.3919
none	0.3629	0.0182	3	0.3494, 0.3506, 0.3887
no_history	0.3569	0.0227	3	0.3375, 0.3444, 0.3887

Table 20: State-encoding ablation at 120,000 episodes, matched to the headline agents' budget. No single block's removal changes performance materially, yet removing almost everything improves it, with no overlap between the seeds of 'minimal' and the full encoding.

The observation was designed to hand the objective to the network explicitly, so the direction of this result runs against the encoding's own rationale. It is also non-monotonic. Removing

the play history alone is the worst option, while removing it along with everything else is the best, which rules out both "more features help" and "fewer features help", and points to an interaction, most plausibly input dimensionality. We report it unresolved.

10.15 What the trained network relies on

Permutation importance complements the ablation: the ablation retrains without a block, this measures what an already-trained policy uses. For a fixed policy π and observation block b, let Π_b replace that block's slice, at every decision, with the same slice taken from an unrelated real game state drawn from a pool of positions. The importance of b is the drop it costs:

$$\Delta_b = w(\pi) - w(\pi \circ \Pi_b). \tag{37}$$

Splicing from a real state rather than zeroing the block is deliberate. Zeroing changes the block's marginal distribution as well as its relationship to the position, and pushes the input off the data manifold, so it can hurt for reasons that have nothing to do with relevance. Π_b preserves the marginal and destroys only the dependence, which is what makes Δ_b readable as reliance.

The current trick and the play history matter most; the explicit ten-status block, the one designed to hand over the objective, is barely used, which is consistent with §10.14 (Supplementary Table S3).

10.16 Hyperparameter sensitivity

Varying learning rate, hidden width and discount around the operating point, one at a time over 7 configurations at 20,000 episodes, moves the win rate against the best heuristic by 0.0637 from best to worst. That spread is smaller than the gap between the learners and the tuned heuristic, so the headline shortfall in §10.1 is not an artefact of a poorly chosen setting; it is also larger than the across-seed standard deviation, so the choice is not free either. Learning rate matters most and the discount least (Supplementary Table S1).

10.17 How much of the game is luck

For TensThenTricks against Random the deal itself explains 0.391 of outcome variance and the seat 0.000; across the pairings decomposed in Table 21 the deal's share runs from 0.290 to 0.391. Conditioning on properties of the deal fixed before play, over 1,200 deals against TensThenTricks, trump length predicts the outcome far more strongly than the number of tens dealt, in a game where only tens score. A team holding eight or more trumps wins 0.785 of hands (1,276 hands) against 0.627 for a team dealt all four tens (284 hands), and the win rate moves 0.592 across the trump-length buckets against 0.254 across the tens-dealt buckets. Long trumps confer control over which side captures the tens, and that control matters more than starting possession of them.

Agent A	Agent B	Deal	Seat	Residual	Mean
TensThenTricks	Random	0.391	0.000	0.609	0.7417
TensThenTricks	TenAware	0.290	0.002	0.708	0.5208

Table 21: Variance decomposition of the outcome. 'Deal' is the cards themselves, i.e. luck; 'seat' is structural seat advantage, which duplicate dealing removes; 'residual' is how the matchup played out.

10.18 Human evaluation

10 people each played 20 hands against ISMCTS250, the strongest agent in Table 7, for 200 hands in total. Participants came from 8 districts of the Terai-Madhesh (Bara, Dhanusha, Mahottari, Nawalparasi, Rautahat, Saptari, Sarlahi and Sunsari), 5 of which also appear in the elicitation cohort of §4. Every participant was between 18 and 40; the enrolment form offered

Condition	Value	Win rate	Hands
tens_dealt	0	0.373	284
tens_dealt	1	0.433	1,182
tens_dealt	2	0.500	1,868
tens_dealt	3	0.567	1,182
tens_dealt	4	0.627	284
trump_length	2	0.000	24
trump_length	3	0.061	98
trump_length	4	0.193	342
trump_length	5	0.249	812
trump_length	6	0.454	1,124
trump_length	7	0.546	1,124
trump_length	8+	0.785	1,276
holds_trump	False	0.499	2,400
holds_trump	True	0.501	2,400
decided_by	tens	0.500	3,092
decided_by	tiebreak	0.500	1,708

Table 22: Outcomes conditioned on properties of the deal fixed before play, for TensThenTricks in self-play. Trump length counts the team's trumps between them and its top bucket is eight or more. A low win rate in a bucket is a property of the situation, not of how it was played.

overlapping bands, so the selections recorded in Table 23 are the options people chose and not a partition of the cohort.

The instrument imposed that age distribution; we did not choose it, and no applicant was refused on grounds of age. Eligibility required knowing the game and playing it unaided on a personal smartphone, and the second condition binds. The most experienced players in these districts are typically 60 to 80, and in this community they are largely not smartphone users: the devices are unfamiliar rather than unavailable, and some of these players did not attend school, so an interface presupposing literacy and touchscreen fluency excludes them. The realised sampling frame is thus device ownership, which here selects for youth.

Of the 11 people who enrolled, 10 finished all their hands. The remaining session stopped after 2 hands, and we drop it from every number that follows, so that a part-played session cannot shift a headline in either direction. Table 23 gives the cohort and Table 24 the per-participant outcomes.

Quantity	Participants
Enrolled	11
Completed all hands	10
Districts represented	8
Bara	1
Dhanusha	1
Mahottari	2
Nawalparasi	1
Rautahat	1
Saptari	1
Sarlahi	2
Sunsari	1
Selected age band 18–25	1
Selected age band 20–30	7
Selected age band 25–35	1
Selected age band 30–40	1

Table 23: Human trial cohort. Counts only; the handling of identifying information is described in the text.

The enrolment form offered overlapping age bands, so the band rows are the options participants selected and not a partition of the cohort. Every participant fell between 18 and 40.

#	Hands won	Rate	Bot on same deals	Difference
1	14/20	0.70	0.72	-0.02
2	12/20	0.60	0.62	-0.02
3	12/20	0.60	0.63	-0.03
4	12/20	0.60	0.54	+0.06
5	10/20	0.50	0.65	-0.15
6	9/20	0.45	0.60	-0.15
7	9/20	0.45	0.62	-0.17
8	9/20	0.45	0.65	-0.20
9	9/20	0.45	0.50	-0.05
10	8/20	0.40	0.60	-0.20

Table 24: Human trial: 10 participants, 200 hands against ISMCTS250. Deals are not common across participants: the dealer of hand $n + 1$ depends on which team lost hand n, so sequences diverge after the first hand.

Participants are numbered, not named; the identity mapping is held in a private consent register and released to editors or reviewers on request. The bot column replays that participant's own deals with a stand-in agent in their seat, averaged over 5 stand-in seeds, so deal luck cancels within each row. For reference, the same agent in all four seats on fresh random deals takes 0.5300 of 200 hands (95% CI [0.4350, 0.6250], containing 0.5), so the seating carries no tilt that would account for the paired gap. In that same control, with identical agents throughout, individual 20-hand rates still ranged from 0.30 to 0.80.

Consent was given before play, in every case. What the interface stores is a participant identifier and the cards played. It stores nothing further: names, phone numbers and districts sit in a separate register, which is not published, and which goes to an editor or reviewer on request so that consent can be checked.

Unpaired comparison. Participants won 0.5200 of hands and the agent 0.4800 (95% CI [0.4200, 0.5350]). The interval resamples participants rather than hands, because hands played by one person share that person's skill, their learning inside the session and their fatigue; treating 200 hands as 200 independent observations would overstate the precision by a wide margin. Individual rates ran from 0.40 to 0.70. The interval contains 0.5, and the smallest effect this sample could detect at 80% power is 0.099 above chance, so read on its own this comparison is simply inconclusive.

This comparison is also the wrong one. The deal explains between 0.290 and 0.391 of outcome variance depending on the pairing, so a win rate measured on one set of cards and compared against 0.5 is mostly measuring which cards turned up. The study was designed to remove that by giving every participant the same deals from a fixed seed (4242), and it does not succeed, for a reason that lies in the game rather than in the software: the dealer of hand $n + 1$ is drawn from the team that lost hand n (R1), so the deal sequence is a function of the participant's own results and diverges after the first hand. Only hand 1 is common to all 10 participants. An analysis assuming a shared deal set across participants would be wrong here.

Paired replay. The comparison is recovered a different way. Each participant's session is reconstructed exactly, by consuming the server's random stream in the original order and dealing against the dealer recorded in the log, then verified card for card against that log; all 10 sessions reproduce exactly, and the analysis refuses to report a baseline if any one of them does not. A stand-in agent is then placed in the participant's own seat and plays their cards against the same opponents, averaged over 5 stand-in seeds. Cards and dealers are held fixed and only the occupant of the seat changes, so deal luck cancels inside each pair rather than being averaged over.

Formally: for participant p let h_p be their win rate over their own 20 hands and b_p the stand-in's rate on the *same* deals from the same seat, averaged over $S = 5$ stand-in seeds. The

estimand is the mean within-pair difference,

$$\hat{\Delta} = \frac{1}{P}\sum_{p=1}^{P}\big(h_p - b_p\big), \qquad b_p = \frac{1}{S}\sum_{k=1}^{S} w\big(\text{stand-in}_k \mid \mathcal{D}_p\big), \tag{38}$$

with the interval from a cluster bootstrap over participants, $B = 20{,}000$ replicates.

Why this recovers what the unpaired comparison could not is a variance argument, and (20) supplies the magnitude. Decompose a single session's outcome into a deal component, a skill component and noise. The unpaired contrast compares h_p against a fixed 0.5, so its sampling variance carries the deal term in full:

$$\mathrm{Var}\big(\hat{\Delta}_{\text{unpaired}}\big) \;\propto\; \sigma^2_{\text{deal}} + \sigma^2_{\text{skill}} + \sigma^2_{\varepsilon}, \qquad \mathrm{Var}\big(\hat{\Delta}_{\text{paired}}\big) \;\propto\; \sigma^2_{\text{skill}} + \sigma^2_{\varepsilon}, \tag{39}$$

since (38) differences that term away within each pair. By (20) it is $\eta^2_{\text{deal}} \in [0.290, 0.391]$, between a third and two-fifths of all outcome variance. Eliminating the dominant variance component, rather than recruiting against it, is what turns these same 200 hands from inconclusive into significant.

The stand-in took 0.6130 of the same deals against the participants' 0.5200. The paired difference is -0.0930 (95% CI [-0.1450, -0.0380]), which excludes zero. On the hands they played, the participants won roughly 9 fewer hands in every hundred than a machine holding their cards would have.

The obvious objection is that the participants' seat is simply a worse seat, or that the loser-deals rotation tilts against whoever occupies it. A control settles both. Running the same agent in all four seats on fresh random deals, under that same rotation, the seat in question takes 0.5300 of 200 hands (95% CI [0.4350, 0.6250]). The interval contains 0.5. No tilt of the size needed to manufacture a -0.0930 gap is present in either the seating or the rotation.

The same control explains why the unpaired numbers came out so weak. Put identical agents in all four seats and the cards become the only thing separating one session from the next, and twenty-hand rates still ranged from 0.30 to 0.80. Twenty hands is close to pure noise no matter who plays them. The participants' spread, 0.40 to 0.70, falls comfortably inside that. So we draw no conclusion at all about how individual participants compare with each other, and Table 24 should not be read as a ranking.

We report the paired figure as the result because the unpaired one points the other way, not merely less firmly. What survives of the round-robin claim is narrower than it first appeared: on identical cards the strongest agent is measurably better than these players, and its raw win rate against them is a coin.

Scope and limitations. Against real players the protocol held up in every part: the interface, the card-for-card reconstruction, the anonymous logging, the paired replay. Three limits are left, and none of them is a software problem.

First, the result describes these 10 people rather than Dahal Jeet players in general, and the shortfall is systematic, not random: by the selection described at the head of this section, the cohort excludes the game's strongest players. The comparison therefore runs against competent but not expert opposition, biasing the deficit upward by an unknown amount, so -0.0930 is a ceiling on the human side rather than a measurement of it at its best.

Reaching those players needs a different instrument, not more recruiting. Assisted sessions, in which the participant plays with physical cards and a researcher enters the moves, would remove the device requirement and are the obvious next step; they introduce a mediator between the player and the interface, which is a confound of its own and would have to be reported as one.

Second, the replay inherits a deal sequence endogenous to the participant's own results, because the losing team deals next. The pairing survives this intact, but the deals are not a fresh sample of the game, so -0.0930 reads "on the hands these players actually faced", not "on hands in general".

Third, the unpaired comparison stays underpowered, and recruitment is the expensive remedy: hands, not participants, bind the precision, so halving the 0.099 bound needs roughly four times the hands. Recruitment continues, and a single script regenerates every number here from the logs.

11 Discussion

Search beats learning on this game, and a tuned heuristic beats every network we trained. The likely reason is (9): with $\rho_{\text{score}} = 4/52$, a whole hand of thirteen tricks yields at most four events that bear on the objective. Search can plan directly toward those events, because (6) is cheap to evaluate at a leaf. A learning agent has to discover them through a signal that is almost everywhere zero.

The encoding result in §10.14 runs against the intent of the design. The observation was built to hand the objective to the network explicitly; measured, that structure does not pay for itself.

11.1 Generality and transfer

The game is regional. Three things about it are not, and this section separates what we measured from what we expect. Everything below is a hypothesis or a methodological recommendation. None of it is a cross-game measurement, because we did not make one; §12 records that as a limitation rather than leaving it implied.

A lexicographic objective, not an additive score. Remark 1 sets (6) against the additive scoring of every benchmark trick-taking game: where those permit a policy to trade a loss on one dimension for a larger gain on another and optimise one scalar, this objective does not. Measured, the constraint bites: Figure 2 shows the two channels pulling in opposite directions hard enough to cancel: 0.6460 on the tens channel against 0.1922 on the tiebreak, netting 0.5024.

An agent facing (6) must therefore decide, at every trick, which of two partially incompatible objectives it is currently playing for, and (24) says the answer depends on a state variable, the tens split, that it does not control. That is a structural problem, not a local one, and Dahal Jeet is a small and cheap instance of it: thirteen tricks, a mean branching factor of 3.880, and an exact endgame solver. We offer it as a testbed for lexicographic multi-objective credit assignment under imperfect information, which is a class of problem rather than a game.

A scoring-density hypothesis that the family can falsify. Our principal negative result, that search beats the tuned heuristic and the tuned heuristic beats every learner, is a fact about this game. The mechanism we propose for it, ρ_{score} in (9), is not specific to this game, and it yields a prediction: *the shortfall of reinforcement learning relative to determinized search should shrink as* ρ_{score} *rises.*

The prediction is testable without inventing a game, because the family supplies the dial and the rules are already documented. Mendikot and Dahal Jeet score four cards in fifty-two. Court Piece and Rang score every trick, $\rho_{\text{score}} = 1$, on a near-identical skeleton (same deck, same partnerships, same follow-suit rule, same thirteen tricks), so the comparison isolates scoring density rather than confounding it with the action space or the trick mechanics. Dehla Pakad varies something else again: its deferred capture releases the pot only when one player takes two consecutive tricks, which lengthens the credit-assignment horizon while leaving ρ_{score} fixed. Three points, one skeleton, two independent axes. We have measured one of them.

Paired replay against human opponents. The result of §10.18 is the one we would most expect to transfer, and it is methodological rather than about Dahal Jeet: on a single set of 200 hands the unpaired and paired analyses are not merely differently precise but discordant, the

former indistinguishable from chance and the latter excluding zero. Read alone, the unpaired number supports the opposite conclusion.

The reason is (39) and has nothing to do with cards: whenever chance contributes a large share of outcome variance, a win rate measured on one realisation of that chance and compared against a fixed reference is mostly measuring the chance. The remedy generalises with the argument. Replay each participant's own realised chance outcomes with a stand-in agent in their position, hold everything but the occupant fixed, and pair. Any game with a large exogenous variance component admits this (backgammon, rummy, poker, most commercial card games), and the cost is a session log detailed enough to reconstruct, which §10.18 shows is achievable and verifiable card by card. We recommend it as the default protocol for human-versus-agent evaluation in stochastic games, and we recommend reporting η^2_{deal} from (20) alongside any such comparison, so that readers can see how much of the measurement was luck before they read the win rate.

Two narrower transfers. Void-constrained determinization (23) applies to any follow-suit game, and our result there is useful precisely because it is negative: the waste is real and large, rising from 1.4% to 64% across a hand, yet at matched wall-clock only one of eight arms survives correction. An engine author working on bridge, skat or hearts can take that as a reason to measure before paying for the constraint. Separately, the pipeline this paper follows (insider elicitation, numbered individually-testable rules, an executable reference engine, then a benchmark) transfers to any traditional game lacking a written specification; the portable parts are the positionality statement, the consent register and the bit-identical reproduction check of §4 and §6. Prior work on Baghchal and on serious games for cultural heritage (Agarwal and Iida, 2018; Raut et al., 2026; Anderson et al., 2010) establishes the frame; our addition is that documentation and benchmark can be a single artifact.

12 Limitations

(a) The optimality bound of (33) is an *endgame* bound. Double-dummy cost grows by roughly a factor of ten per additional trick, which puts whole-hand solving out of reach at any budget we could justify.

(b) Neither game-theoretic baseline carries its usual guarantee. Deep CFR uses depth-limited external sampling with rollout evaluation, which has no equilibrium guarantee in a four-player partnership game, and is reported as a baseline rather than a solution. Exploitability is a lower bound only, for the reason set out in Remark 4: the oracle class must contain a genuine best response, and here it does not.

(c) The encoding inversion is unexplained.

(d) One ruleset, one region, one language community; the provenance is retrospective and without contemporaneous field notes.

(e) The human trial is limited in three ways, and recruitment is continuing. Its 10 participants suffice to exclude a large agent advantage but not to resolve a small one in either direction. The smartphone requirement excludes the 60-to-80 cohort who play the game most and supplied much of the ruleset, so the measured human side is competent but not the strongest available and the paired deficit is an upper bound on it; assisted play with physical cards is the remedy, at the cost of a mediator. And the deals are not common across participants, the paired replay on each participant's own cards controlling for deal luck instead (§10.18).

(f) The transfer claims of §11.1 are hypotheses supported by this game's structure, not measurements across games: the scoring-density prediction is untested at a single measured point, and the paired-replay recommendation, though it rests on a general variance argument, has been demonstrated on this game alone.

13 Compute

All experiments ran on a single laptop CPU; Table 25 records why the integrated GPU backend was benchmarked and rejected. Profiling attributes half of training wall-clock to Python-to-tensor conversion rather than to matrix arithmetic, which is the underlying reason the accelerator did not pay.

Quantity	Value
Processor	M1 Pro
Logical cores	10
Memory	16.0 GiB
Accelerator	none; CPU only
Result artifacts	81
Measured CPU-hours	134.9

Table 25: Compute and hardware. Every experiment ran on the CPU of a single laptop; no GPU acceleration was used.

The platform's integrated GPU backend was benchmarked and rejected: at batch size 1, which dominates reinforcement-learning rollouts, it is roughly four times slower than the CPU for networks of this size.

14 Conclusion

Dahal Jeet had no published rule description. This paper supplies one: a numbered, individually testable specification with documented provenance across seven districts. It then uses that specification to ask what the game is worth as a computational subject. The answer is that it is strategically non-trivial and that its difficulty has a nameable source. By (9) only four of fifty-two cards score, so a full hand yields at most four events bearing on (6), and that objective is lexicographic rather than additive. Determinized search handles this structure; every learning method we trained does not. Across 325 pairings under one protocol, ISMCTS leads at Elo 1663.2 and no network reaches a metaheuristically tuned heuristic.

Three results generalise past the game. A minimal state encoding beats a rich hand-designed one by +0.0644 with no seed overlap, against the intent of the design. Every deterministic heuristic tested is exploitable by a search that knows its policy, while random play is not. And on 200 hands against 10 players, the unpaired and paired analyses disagree in direction rather than only in precision, because the deal contributes between 0.290 and 0.391 of outcome variance; replaying each participant's own cards with a stand-in agent is what recovers the comparison. We recommend that protocol wherever chance dominates outcome variance, and we offer the game itself as a compact testbed for lexicographic objectives under imperfect information.

Statements and Declarations

Funding. The authors received no funding for this work.

Competing interests. The authors declare no competing interests relevant to the content of this article.

Data and code availability. The complete code base, all trained models, and every result artifact, including the per-deal outcome vectors for all 325 tournament pairings, will be made available to reviewers and editors on request, and released publicly on publication. Every statistic reported here can be recomputed from those files by a single script, and the identities asserted by the equations are rechecked against the artifacts by a second. The 81 artifacts and 134.9 CPU-hours, all on a single M1 Pro laptop, are itemised in Table 25.

Ethics and responsible use. Every participant in the human trial gave informed consent before playing. The interface records only a participant identifier and the cards played; names, phone numbers and districts are held in a separate register that is not published and is released to an editor or reviewer on request so that consent can be verified. Participants are numbered rather than named throughout, and the elicitation cohort is attributed at district level only, at the players' own preference.

Declaration of generative AI and AI-assisted technologies. During the preparation of this work, the authors used Claude (Anthropic) in order to improve the grammar, punctuation, and readability of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication. No generative AI tool was used to generate the scientific content, experimental design, results, data analysis, or conclusions presented in this work. All references cited have been verified by the authors against their original sources.

Author contributions. **Raju Kumar Yadav:** Conceptualization, methodology, software development, data curation, investigation, experimentation, formal analysis, visualization, preparation of figures and tables, and writing of the original manuscript. **Ganesh Gautam:** Supervision, conceptual guidance, methodology, validation of the research approach, interpretation of results, project administration, and critical review, revision, and editing of the manuscript. All authors contributed to the discussion and interpretation of the findings, reviewed the manuscript critically for important intellectual content, and read and approved the final version of the manuscript.

Acknowledgements

Thanks to the many anonymous players across Sarlahi, Dhanusha, Mahottari, Siraha, Saptari, Parsa and Rautahat who taught, played and corrected this game over the years, and to the 10 participants of the human trial. They are unnamed here by their own preference.

References

Sakshi Agarwal and Hiroyuki Iida. Analyzing thousand years old game tigers and goats is still alive. *Asia-Pacific Journal of Information Technology & Multimedia*, 07:1–11, 12 2018. doi: 10.17576/apjitm-2018-0702-01.

Asaf Amit and Shaul Markovitch. Learning to bid in bridge. *Machine Learning*, 63(3):287–327, 2006. doi: 10.1007/s10994-006-6225-2.

Eike Falk Anderson, Leigh Mcloughlin, Fotis Liarokapis, Christopher Peters, Panagiotis Petridis, and Sara Freitas. Developing serious games for cultural heritage: a state-of-the-art review. *Virtual Real.*, 14(4):255–275, December 2010. ISSN 1359-4338.

Tathagata Banerjee, Anushka De, Subhamoy Maitra, and Diganta Mukherjee. Skill vs. chance quantification for popular card & board games, 2025. URL `https://arxiv.org/abs/2410.14363`.

Alexandre Bayen, Jiaxuan Gao, Akash Velu, Eugene Vinitsky, Yu Wang, Yi Wu, and Chao Yu. The surprising effectiveness of ppo in cooperative multi-agent games, 01 2022.

Noam Brown, Adam Lerer, Sam Gross, and Tuomas Sandholm. Deep counterfactual regret minimization, 2019. URL `https://arxiv.org/abs/1811.00164`.

Gal Cohensius, Reshef Meir, Nadav Oved, and Roni Stern. Bidding in spades, 2020. URL `https://arxiv.org/abs/1912.11323`.

Peter Cowling, Edward Powley, and Daniel Whitehouse. Information set monte carlo tree search. *IEEE Transactions on Computational Intelligence and AI in Games*, 4:120–143, 06 2012a. doi: 10.1109/TCIAIG.2012.2200894.

Peter I. Cowling, Colin D. Ward, and Edward J. Powley. Ensemble determinization in monte carlo tree search for the imperfect information card game magic: The gathering. *IEEE Transactions on Computational Intelligence and AI in Games*, 4(4):241–257, 2012b. doi: 10.1109/TCIAIG.2012.2204883.

M. L. Ginsberg. Gib: Imperfect information in a computationally challenging game. *Journal of Artificial Intelligence Research*, 14:303–358, June 2001. ISSN 1076-9757. doi: 10.1613/jair.820. URL `http://dx.doi.org/10.1613/jair.820`.

Johannes Heinrich and David Silver. Deep reinforcement learning from self-play in imperfect-information games, 2016. URL `https://arxiv.org/abs/1603.01121`.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL `https://arxiv.org/abs/1503.02531`.

Aadi Juthani. Optimizing LLM strategies for playing Mendikot using prompt engineering. *International Journal for Multidisciplinary Research*, 6(6), 2024. ISSN 2582-2160. doi: 10.36948/ijfmr.2024.v06i06.30130.

Sahaj Raj Malla. Heuristic optimization of AI agents for Dhumbal. *Scientific Reports*, 2026. doi: 10.1038/s41598-026-68206-1.

John McLeod. 3-2-5 (Teen Do Paanch). Pagat.com, Rules of Card Games, n.d.a. URL `https://www.pagat.com/whist/3-2-5.html`.

John McLeod. Call Bridge (Call Break). Pagat.com, Rules of Card Games, n.d.b. URL `https://www.pagat.com/auctionwhist/call_bridge.html`.

John McLeod. Court Piece, Rang and Double Sir. Pagat.com, Rules of Card Games, n.d.c. URL `https://www.pagat.com/whist/rang.html`.

John McLeod. Dehla Pakad. Pagat.com, Rules of Card Games, n.d.d. URL `https://www.pagat.com/pointtrk/dehlapakad.html`.

John McLeod. Marriage. Pagat.com, Rules of Card Games, n.d.e. URL `https://www.pagat.com/rummy/marriage.html`.

John McLeod. Mendikot. Pagat.com, Rules of Card Games, n.d.f. URL `https://www.pagat.com/pointtrk/mendikot.html`.

John McLeod. Card Games of Nepal. Pagat.com, National Card Games, n.d.g. URL `https://www.pagat.com/national/nepal.html`.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–33, 2015. doi: 10.1038/nature14236.

Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning, 2016. URL `https://arxiv.org/abs/1602.01783`.

Andrew Y. Ng, Daishi Harada, and Stuart J. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, page 278–287, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1558606122.

Ranjit Raut, Aarav Subedi, Sagun Rai, Aaryan Shakya, and Manoj Shakya. Deep reinforcement learning to master the asymmetric strategy of baghchal, 2026. URL `https://arxiv.org/abs/2607.18296`.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL `https://arxiv.org/abs/1707.06347`.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017. doi: 10.1038/nature24270.

Christopher Solinas, Douglas Rebstock, and Michael Buro. Improving search with supervised learning in trick-based card games. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):1158–1165, July 2019. ISSN 2159-5399. doi: 10.1609/aaai.v33i01.33011158. URL `http://dx.doi.org/10.1609/aaai.v33i01.33011158`.

Hado van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning, 2015. URL `https://arxiv.org/abs/1509.06461`.

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado van Hasselt, Marc Lanctot, and Nando de Freitas. Dueling network architectures for deep reinforcement learning, 2016. URL `https://arxiv.org/abs/1511.06581`.

Wikipedia contributors. Court piece. Wikipedia, n.d. URL `https://en.wikipedia.org/wiki/Court_piece`. Carries a "needs more citations" banner and unsourced passages; used as a lead, not as an authority.

Daochen Zha, Kwei-Herng Lai, Yuanpu Cao, Songyi Huang, Ruzhe Wei, Junyu Guo, and Xia Hu. Rlcard: A toolkit for reinforcement learning in card games, 2020. URL `https://arxiv.org/abs/1910.04376`.

Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, NIPS'07, page 1729–1736, Red Hook, NY, USA, 2007. Curran Associates Inc. ISBN 9781605603520.